Community Detection in Social Networks: Multilayer Networks and Pairwise Covariates

Sihan Huang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2020

Abstract

Community Detection in Social Networks: Multilayer Networks and Pairwise Covariates

Sihan Huang

Community detection is one of the most fundamental problems in network study. The stochastic block model (SBM) is arguably the most studied model for network data with different estimation methods developed with their community detection consistency results unveiled. Due to its stringent assumptions, SBM may not be suitable for many real-world problems. In this thesis, we present two approaches that incorporate extra information compared with vanilla SBM to help improve community detection performance and be suitable for applications.

One approach is to stack multilayer networks that are composed of multiple single-layer networks with common community structure. Numerous methods have been proposed based on spectral clustering, but most rely on optimizing an objective function while the associated theoretical properties remain to be largely unexplored. We focus on the ‘early fusion’ method [114], of which the target is to minimize the spectral clustering error of the weighted adjacency matrix (WAM). We derive the optimal weights by studying the asymptotic behavior of eigenvalues and eigenvectors of the WAM. We show that the eigenvector of WAM converges to a normal distribution as in [129], and the clustering error is monotonically decreasing with the eigenvalue gap. This fact reveals the intrinsic link between eigenvalues and eigenvectors, and thus the algorithm will minimize the clustering error by maximizing the eigenvalue gap. The numerical study shows that our algorithm outperforms other state-of-art methods significantly, especially when signal-to-noise ratios of layers vary widely. Our algorithm also yields higher accuracy result for S&P 1500 stocks dataset than competing models.

The other approach we propose is to consider heterogeneous connection probabilities to remove the strong assumption that all nodes in the same community are stochastically equivalent, which may not be suitable for practical applications. We introduce a pairwise covariates-adjusted stochastic block model (PCABM), a generalization of SBM that incorporates pairwise covariates information. We study the maximum likelihood estimates of the coefficients for the covariates as well as the community assignments. It is shown that both the coefficient estimates of the covariates and the community assignments are consistent under suitable sparsity conditions. Spectral clustering with adjustment (SCWA) is introduced to fit PCABM efficiently. Under certain conditions, we derive the error bound of community estimation under SCWA and show that it is community detection consistent. PCABM compares favorably with the SBM or degree-corrected stochastic block model under a wide range of simulated and real networks when covariate information is accessible.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Foremost I wish to express my sincere appreciation to my advisor, Professor Yang Feng, for his continuous guidance and encouragement. He taught me to be professional and find the right direction when I am lost. Without his persistent help, this thesis would not have been possible.

I would like to thank Professor Haolei Weng for his suggestions for the second chapter. Also, the serving of Professor Zhiliang Ying, Professor Tian Zheng, and Professor Daniel Hsu on my committee is truly appreciated. Their insightful comments are valuable to improve my thesis.

I have been blessed with an accompany of numerous friends who gave me so much support and joy. There are too many to thank by name, but I am especially grateful to Yang Kang, Liwei Wu, Fan Shi, Yi Hong, Zhi Li, Yuanzhe Xu, Boyuan Zhang, Chaoyu Yuan and Jing Wu, whose friendship has been so important to me throughout the hard periods of my life.

Finally, I wish to acknowledge the support and great love of my family: my parents Jian Huang and Min Zhang, for giving birth to me and supporting me spiritually throughout my life.

To my beloved parents who always support me to explore the possibilities of my life.

Chapter 1: Introduction and Background

The network describes the connections among subjects in a population of interest. Its wide applications have attracted researchers from different fields, which includes but not limited to social sciences [135, 93], physics [10], computer science [110], biology [79] and so on. Among all problems for network, community detection is one of the most studied ones, and the stochastic block model (SBM) has long been used as a canonical model for that. Although theoretical properties of classical SBM have been studied thoroughly, the theoretical framework hasn't been established for its variants. We're interested in the expanded version of SBM, say when there is extract information beside the network edges, how we could exploit that information and make a better inference.

We will review the history of network analysis in Section 1.1, especially focusing on the researches of community detection and SBM. In Section 1.2, we provide an overview of our two variants of SBM. Section 1.3 introduces some basic notations that will be used later.

1.1 Literature review

The very first researches for network date back to the middle of last century, marked by some empirical studies in social psychology and sociology [104, 124, 103, 130]. The works in social science ignite researchers' passion for mathematical insights of network models. Erdős-Rényi model, which became the canonical model in network analysis later, was first introduced in the 1960s [55]. The random graph theory that later arose was mostly based on that. People studied the asymptotic behavior of the Erdős-Rényi model, which has n nodes and each edge is generated by the homogeneous probability p. The 'phase transition' that later became well-known associated with the value of $pn = 1$, at which point many small clusters will merge into a single giant connected component.

Erdős-Rényi model was extended and generalized in the 1980s. The models have the most lasting impact include but are not limited to p_1 model [74, 60], multidimensional network structures [59], SBM [73], exponential random graph models [64, 126, 136, 113, 118] and so on. More literature in the 1990s linked random graph models with stochastic processes and dig deeper into the probabilistic properties [43, 52]. The machine learning approach is what emerged at the turn of this century. Some empirical studies were done by [57, 83, 82, 84]. More recent efforts focus on the variants of SBM. [72] combines SBM with latent space models. [70] introduced model-based clustering. [8] includes the latent Dirichlet allocation [28] to provide a novel generative model, which is a kind of soft clustering [56].

Communities in a network can be intuitively understood as groups of nodes that are densely connected within groups while sparsely connected between groups. The goal of community detection is to partition the nodes into different sets. Not only does identifying network communities help better understand the structural features of the network, but it also offers practical benefits. Besides its original target to understand sociological behavior [67, 63], other applications of community detection have also been widely explored. Biologists classify protein function from the protein-protein interaction network [7, 9, 38, 100]. By analyzing communication networks [18], people can detect possible latent terrorist cells. With the emerging of online 'networking communities' such as *Facebook, Twitter* and *Instagram*, the need for a recommender system based on the similarity of nodes within the network also becomes nonnegligible [71, 94, 120, 140]. Other applications include but are not limited to gene expressions [45, 77], natural language processing [19], image segmentation [123] and more.

Classical community detection methods fall into three categories. The first category is algorithm-based. One example is hierarchical clustering, which gradually removes edges until disconnected clusters are obtained. More reviews can be found in [107]. The second category is criterion-based. Such methods optimize some criterion to obtain the best partition of the network. This criterion often refers to some cuts or modularities, which include Newman-Girvan modularity [109], the ratio cut [137], the normalized cut [123] and more. The third category is model-based, which is the main focus of this thesis. This kind of model builds a generative probabilistic model for the network and tries to identify the latent community labels by fitting the model. Besides SBM [125, 112] and mixed membership SBM that have been mentioned, other examples include degree-corrected stochastic block models (DCBM) [80, 147], the latent position cluster model [70], latent space model [72], etc. One thing that needs to remember is that those categories mentioned above are just rough philosophical descriptions for different methods rather than strict divisions. For example, we'll see that the model in Chapter 3 can be both model-based and criterion-based, just like [147, 24].

Among all the models, SBM might be the most natural one for describing the community detection problem because of its neat form. Although the block structure of the network has already been discussed as early as the 1970s [95], the random graph framework only came up a decade later in [73], after numerous papers for deterministic blocks arose [14, 51, 50]. People in different fields name it differently, for example, planted partition model (PPM) in theoretical computer science [33, 53, 30], inhomogeneous random graph in math [29] and so on.

The enthusiasm for SBM theories mainly focuses on the consistency of community estimation. Analogous to the concept of random variable convergence, for community detection, we have strong consistency and weak consistency, which we will define in detail later. [49] pictured the weak recovery in detail and conjectured that there exist phase transition phenomena for weak consistency at the Kesten-Stigum threshold. The later activation of establishing phase transitions should be credited to this paper. [106] proved the impossibility part of the conjecture for two symmetric communities, while the positive part for two communities was shown in [101]. For non-symmetric case, [31] proved that the threshold can be achieved under a more general setting which covers asymmetric SBM of two communities. This borrows the idea from the 'spectral redemption' paper [85], which introduces the nonbacktracking operator. More general cases for arbitrarily many communities were shown in [3, 2], using a high-order nonbacktracking matrix and a spectral-message passing. Crossing information-theoretic Kesten-Stigum threshold under more communities were further discussed in [4, 17]. For exact recovery, the phase transition also exists, though in the logarithmic rather than constant degree regime. For years, researchers worked on this problem [33, 53, 30, 125, 46, 102, 24, 42, 133] until the sharp threshold was given in [1, 105] for two symmetric communities. For general SBM, the parallel results were shown in [3, 5].

To efficiently perform inference for SBM under different settings, people designed various algorithms. Some popular ones include modularity maximization [109], likelihood methods [24, 42, 147, 12, 35], spectral clustering [16, 119, 61, 78, 91, 122], variational inference [37, 26, 48, 35], belief propagation [49], convex optimization [41], semidefinite programming [1, 34, 108], penalized local maximum likelihood estimation [65] and more.

Among all of them, spectral clustering stands out for its neat form and simplicity to implement, so it's long been a fascination for researchers. The detailed tutorial of spectral clustering can be found in [132], and we just list a few important literature here. [119] proved the consistency for spectral clustering with the normalized graph Laplacian, but it assumes the average degree grows faster than $\log n$. The same assumption holds in [78] and [122]. The former one proposed spectral clustering on ratios-of-eigenvectors (SCORE) for the DCBM. The latter one compared normalized and unnormalized spectral clustering. Shortly after, [91] proved consistency when the maximum expected node degree is of order $\log n$ or higher. Also, an approximate k-means algorithm solvable in polynomial time [86] is considered in [91] rather than the global optimizer of k-means. However, on sparser graphs (e.g., average degree is of order $O(1)$), spectral clustering on standard graph Laplacian performs poorly [37, 34, 13, 89], that's where the regularized graph Laplacian came into place [37]. It's proven to misclassify at most arbitrarily small proportion of nodes [88]. There are some other variants of spectral clustering for different purposes. [36] introduced the notion of a degree-corrected graph Laplacian for the extended PPM. [61] required only an upper bound of the number of communities for the modified adjacency matrix to guarantee the consistency of spectral clustering. [117] applied the regularized graph Laplacian to the canonical spectral clustering algorithm and provided the error bound for DCBM and extended PPM.

Besides what we've mentioned above, there are some other research topics on SBM's variants over the recent years. One concern is about the 'outliers' in the network, which refers to the node that doesn't belong to any group. Similar to regression, robust community detection algorithm is proposed to solve this problem. [146] extracts one community at a time, allowing for arbitrary structure in the remainder of the network, which can include weakly connected nodes. [34] extended SBM to generalized SBM, allowing the 'outliers' to connect with other nodes arbitrarily, and the model is fitted by SDP. Also, some people work on local community detection [62, 44, 141, 131, 116], the target of which is to partition some nodes of interests rather than all. Another topic is to determine the number of communities since most researches we have mentioned assume the number of community to be known, which is unrealistic. [146] sequentially extracts communities until the remaining part behaves like an Erdős-Rényi graph from a hypothesis testing. Two papers designed hypothesis testing based on random matrix theories. One [25] is based on the limiting distribution of the principal eigenvalue of a centered and scaled adjacency matrix, and the other [90] is based on the largest singular value of a residual matrix obtained by subtracting the estimated block mean effect from the adjacency matrix. Some BIC approaches have also been proposed [121, 134, 75]. Another topic is nodal covariates, which we will discuss in more details in Chapter 3.

1.2 Overview of the thesis

In this thesis, we proposed two models that take care of extra information besides the edge connection under the single layer SBM model.

In Chapter 2, we used a weighted method to combine the information in a multilayer PPM. We focus on the 'early fusion' method [114], of which the target is to minimize the spectral clustering error of the weighted adjacency matrix (WAM). It's shown in Section 2.3.1 that the clustering error is monotonically decreasing with a signal-to-noise ratio (SNR) that is determined by the connectivity matrix of the multilayer network. This is proven by studying the asymptotic distribution of eigenvectors, which converges to a normal distribution as [129]. By maximizing the SNR, we derived the optimal weights. We also showed in Section 2.3.2 that the limit of the eigenvalue gap is also a monotonic function of SNR in some scenarios, which allows us to minimize the clustering error by maximizing the eigenvalue gap.

In Chapter 3, we introduced a pairwise covariates-adjusted stochastic block model (PCABM), which is a generalization of SBM that incorporates nodal information. We studied the maximum likelihood estimates of the coefficients for the covariates as well as the community assignments. It is shown in Section 3.3.1 that both the coefficient estimates of the covariates and the community assignments are consistent under mild conditions. As a more efficient algorithm than the likelihood method, a variant of spectral clustering, spectral clustering with adjustment (SCWA), is introduced in Section 3.3.2 to estimate PCABM. We derived the error bound of community estimation by SCWA and showed that it is community detection consistent, which achieves the same convergence rate as in [91].

1.3 Notation and terminology

For the convenience of reference, we define some notations that will be commonly used in both Chapter 2 and 3.

First, we clarify some mathematical terminologies. We define $I_K \in \mathbb{R}^{K\times K}$ to be the identity matrix, $J_K \in \mathbb{R}^{K\times K}$ to be all-one matrix and $\mathbf{1}_K$ to be all-one vector. When there is no confusion, the subscript K will be omitted. For any positive integer K, $[K]$ denotes the set of numbers $1, \cdots, K$. For a vector $\mathbf{x} \in \mathbb{R}^K$, $D(\mathbf{x}) \in \mathbb{R}^{K\times K}$ represents the diagonal matrix whose diagonal elements are the entries of $\mathbf{x}$. For an event A, $\mathbb{1}_A$ denotes the indicator function. For two real number sequences x_n and y_n, we say $x_n = o(y_n)$ if $\lim_{n\to\infty} x_n/y_n = 0$, $x_n = O(y_n)$ if $\limsup_{n\to\infty} |x_n|/y_n \leq \infty$, $x_n = \omega(y_n)$ if $\lim_{n\to\infty} |x_n/y_n| = \infty$ and $x_n = \Theta(y_n)$ if $\exists c_1, c_2, n_0 > 0, \forall n > n_0, c_1 y_n \leq x_n \leq c_2 y_n$.

For a square matrix $X \in \mathbb{R}^{n\times n}$, let $\|X\|$ be the operator norm, $\|X\|_F = \sqrt{\mathrm{trace}(X^T X)}$, $\|X\|_\infty = \max_i \sum_{j=1}^n |X_{ij}|$, $\|X\|_0 = |\{ij | X_{ij} \neq 0\}|$, $\|X\|_{\max} = \max_{ij} |X_{ij}|$ and $\|X\|_1 = \max_{i\in[n]} \sum_{j=1}^n |X_{ij}|$. For index sets $I, J \subset [n]$, $X_{I\cdot}$ and $X_{\cdot J}$ are the sub-matrix of X consisting of the corresponding rows and columns. Similarly, for a vector $\mathbf{x} \in \mathbb{R}^n$, $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$, $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ and $\|\mathbf{x}\|_\infty = \max_i |x_i|$. The Kronecker power is defined as $\mathbf{x}^{\otimes(k+1)} = \mathbf{x}^{\otimes k} \otimes \mathbf{x}$, where $\otimes$ is the Kronecker product.

Now we provide the mathematical form of SBM. Consider a graph with n nodes and K communities, where K is fixed and does not increase with n. We only focus on undirected graph without self-loops, whose all edge information is incorporated into a symmetric adjacency matrix $A = [A_{ij}] \in \mathbb{N}^{n \times n}$ with diagonal elements being zero. The total count of possible edges is $N_n \equiv n(n-1)/2$. The true node labels $\mathbf{c} = (c_1, \cdots, c_n)^T \in [K]^n$ are drawn independently from a multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)^T$, where $\|\boldsymbol{\pi}\|_1 = 1$ and $\pi_k > 0$ for all k. The community detection problem is aiming to find a disjoint partition of the node set, or equivalently, a set of node labels $\mathbf{e} = \{e_1, \cdots, e_n\} \in [K]^n$, that is as close to the true label $\mathbf{c}$ as possible, up to a permutation. In classical SBM, we assume $A_{ij} \sim \text{Bernoulli}(B_{c_i c_j})$, where the connectivity matrix $B = [B_{ab}] \in (0,1)^{K \times K}$ is a symmetric matrix with no identical rows. Usually, we need to consider a sparse setting where the connectivity matrix B scale with n rather than fixed. Assume $B = \rho_n \Omega$ with Ω fixed and $\rho_n \longrightarrow 0$ as $n \longrightarrow \infty$. In this case, $\varphi_n \equiv n\rho_n$ represents the expected degree. n_k denotes the number of nodes in group k, so $\sum_{k=1}^{K} n_k = n$. Also, we sometimes need the probability matrix $P \in (0,1)^{n \times n}$, where $P_{ij} = B_{c_i c_j}$. We say $A \sim \text{SBM}(\Omega, \mathbf{c}, \rho_n)$ if the adjacency matrix $A \in \{0,1\}^{n \times n}$ is symmetric and for all $i < j$, $A_{ij} \sim \text{Bernoulli}(\rho_n \Omega_{c_i c_j})$.

At last, we briefly discuss the classical spectral clustering with k-means for SBM. Define the membership matrix $M \equiv [\mathbb{1}_{c_i = j}] \in \mathcal{M}_{n,K}$, where $\mathcal{M}_{n,K}$ is the space of all $n \times K$ matrices where each row has exactly one 1 and $(K-1)$ 0's. Note that M contains the same information as $\mathbf{c}$ and it's only introduced to facilitate the discussion of spectral methods. It is easy to see that $P = MBM^T$. For adjacency matrix A, we consider only K largest eigenvalues in absolute values $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_K|$ and their corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_K$. Suppose the matrix $\Lambda_A \equiv D((\lambda_1, \cdots, \lambda_K)^T) \in \mathbb{R}^{K \times K}$ and $U_A \equiv [\mathbf{u}_1, \cdots, \mathbf{u}_K] \in \mathbb{R}^{n \times K}$, where $U_A^T U_A = I_K$. Since $\mathbb{E}[P] = A$, U_P is expected to be close to U_A, where U_P comes from the decomposition $P = U_P \Lambda_P U_P^T$. Only K unique rows appear in U_P, which corresponds to different communities, so the k-means clustering on the rows of U_A should lead to a good estimate of M. To evaluate the performance of clustering, we use the measure adjusted rand index (ARI) [76], which is commonly used in clustering problem. We summarize the procedure of spectral clustering algorithm as follows.

Algorithm 1.1 Spectral clustering

```
Input: adjacency matrix $A \in \mathbb{N}^{n \times n}$ and community number $K$
Output: community estimation $\hat{M}$
Consider the decomposition $A = U_A \Lambda_A U_A^T$, which corresponds to $K$ largest eigenvalues.
Treating each of the $n$ rows of $U_A$ as a point in $\mathbb{R}^K$, run $k$-means with $K$ clusters. This creates a $K$ partition of $[n]$, from which we could produce the estimated membership matrix $\hat{M}$.
```

Chapter 2: Optimal Weights for Multilayer Networks

2.1 Introduction

Tons of papers only focus on single-layer network, while multilayer networks are very common in the real world. For example, for professors working in the same department, there might be a network of lunch, a network of research and a network of race. People may have different social patterns in different social media such as *Facebook*, *Twitter* and *Instagram*. Multilayer network displays diverse relations among the same set of nodes, so how to combine the information across different layers turns out to be a very important problem.

Among limited literature using an 'early fusion' technique, most of them assume an SBM framework [73]. [40] proposes to combine multiple layers linearly and it explores the threshold for detection. However, the conclusion is weak because of the dense setting and optimal weights choice isn't discussed. [69] derives the consistency result when the number of layers grows and the number of nodes does not. [114] compares different algorithms for multilayer networks. Some are presented in a matrix factorization manner, while some use the modularity method, but nothing much has been discussed about the mean adjacency matrix.

Our target is to find an adaptive algorithm to find the optimal weights vector to linearly combine adjacency matrices. The 'optimal' is defined as minimizing the spectral clustering error. To achieve that, we studied the limit distribution for eigenvectors and eigenvalues of WAM and derived the optimal weights' formula based on that. We showed that under some sparsity scales, the optimal weights we derived could improve the clustering error rate asymptotically.

The rest of this chapter is organized as follows. Section 2.2 will introduce the intuition of the problem and our approach to solve the problem. Theory and the corresponding algorithms are presented in Section 2.3. Simulation results are described in Section 2.4, and one real example of

S&P 1500 data is presented in Section 2.5. We finally conclude this chapter with a short discussion in Section 2.6. Proofs are summarized in Section 2.7.

2.2 Multilayer network model

In this chapter, we will only consider balanced SBM, which means $\pi_1 = \cdots = \pi_K = 1/K$. PPM, one special case of SBM, will be discussed in this chapter as well. Under PPM, the connectivity matrix has a special form of

$$\Omega = (\Omega_{in} - \Omega_{out})I_K + \Omega_{out}J_K \in (0,1)^{K\times K}.$$

This means the within-class probabilities of PPM are all Ω_{in} while the between-class probabilities are all Ω_{out}. Here, we consider only the assortative network, which requires $\Omega_{in} \geq \Omega_{out}$.

We can simply stack single layer network to define the multilayer network model. A multilayer network of L layers is a collection of L networks that share the same nodes but with different edges. Specifically, a multilayer stochastic block model (MSBM) is a multilayer network that each layer follows a SBM with consensus group assignment $\mathbf{c}$, i.e., $A^{(l)} \sim \mathrm{SBM}(\Omega^{(l)}, \mathbf{c}, \rho_n)$ for $l \in [L]$. We write that as $A^{[L]} \sim \mathrm{MSBM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$ for short. If every layer is PPM, we call it a multilayer planted partition model (MPPM) and write it as $A^{[L]} \sim \mathrm{MPPM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$. Throughout this chapter, we will assume the following sparsity condition.

Condition 2.1. *$n\rho_n \geq C\log n$ for some positive constant $C > 0$ and $\rho_n \to 0$ as $n \to \infty$.*

For a multilayer network, though different layers share a common community structure, the information contained in different layers varies. Some layers may connect more densely within groups compared with others. When aggregating such a multilayer network, the former layer should overweigh the latter. Starting from this intuition, we wish to find the optimal weights in an 'early fusion' style [40, 114] if it ever exists. The target is to find the optimal weights' vector $\mathbf{w} = (w_1, \cdots, w_L)^T \in \mathcal{W}_L$ to minimize the mismatch error of spectral clustering on the WAM $A^{\mathbf{w}} = \sum_{l=1}^{L} w_l A^{(l)}$. Here, $\mathcal{W}_L = \{\mathbf{w} \in \mathbb{R}^L \mid \|\mathbf{w}\|_1 = 1, w_l \geq 0\}$. The mismatch error refers to the proportion of error up to a permutation, which is $n^{-1}\min_{\sigma\in\mathcal{P}_K}\sum_{i=1}^{n}\mathbb{1}_{\sigma(\hat{c}_i)\neq c_i}$. Here, $\mathcal{P}_K$ is the collection of all permutation functions of $[K]$.

2.3 Theory and algorithms for the optimal weights

We proposed two approaches for finding the optimal weights. The first one in Section 2.3.1 is based on a closed-form formula of connectivity matrix, which can be estimated empirically in practice. The second one in Section 2.3.2 is based on the maximization of the eigenvalue gap. The theoretical supports behind those two algorithms are the asymptotic distributions of eigenvectors and eigenvalues of WAM respectively, which are established on some previous results for random matrices.

2.3.1 Closed-form solution for optimal weights

Under the spectral clustering algorithm, if the eigenvector features for different groups are separable, the clustering results will be good. Similar to [129], we studied the eigenvectors' asymptotic distribution of WAM and found the distribution of eigenvectors is closely correlated with the SNR, which is defined as

$$\tau_n^{\mathbf{w}} \equiv \frac{1}{2}\left(\frac{n\rho_n}{K}\right)^{1/2}\frac{\Omega_{in}^{\mathbf{w}}-\Omega_{out}^{\mathbf{w}}}{\{\Omega_{in}^{\mathbf{w}^2}+(K-1)\Omega_{out}^{\mathbf{w}^2}\}^{1/2}},$$

where $\Omega_{in}^{\mathbf{w}}=\sum_{l=1}^{L}w_l\Omega_{in}^{(l)}$, $\Omega_{out}^{\mathbf{w}}=\sum_{l=1}^{L}w_l\Omega_{out}^{(l)}$, $\Omega_{in}^{\mathbf{w}^2}=\sum_{l=1}^{L}w_l^2\Omega_{in}^{(l)}$, and $\Omega_{out}^{\mathbf{w}^2}=\sum_{l=1}^{L}w_l^2\Omega_{out}^{(l)}$. As we can see, the nominator $\Omega_{in}^{\mathbf{w}}-\Omega_{out}^{\mathbf{w}}$ measures the difference between two connection probabilities while the denominator $\{\Omega_{in}^{\mathbf{w}^2}+(K-1)\Omega_{out}^{\mathbf{w}^2}\}^{1/2}$ represents the standard deviation of the connection probability. Thus, $\tau_n^{\mathbf{w}}$ is just a normalized version of the original SNR. We assume the following condition for all theoretical results, which helps us to analyze the rate. The monotonic relation between clustering error and SNR is stated in Proposition 2.1.

Condition 2.2. $\tau_\infty^{\mathbf{w}} \equiv \lim_{n\to\infty}\tau_n^{\mathbf{w}}$ *exists.*

Proposition 2.1. *For balanced $A^{[L]} \sim \mathrm{MPPM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$ under Condition 2.1 and 2.2, if $\mathbf{w} \in \mathcal{W}_L$, the asymptotic Bayes error rate for spectral clustering on WAM $A^{\mathbf{w}}$ is monotonic decreasing with $\tau_\infty^{\mathbf{w}}$. Specifically,*

1. *If $\tau_\infty^{\mathbf{w}} = \infty$, the asymptotic error is $1/K$.*

2. *If $\tau_\infty^{\mathbf{w}} = 0$, the asymptotic error is 1.*

3. *If $0 < \tau_\infty^{\mathbf{w}} < \infty$, the asymptotic error is a constant between $1/K$ and 1.*

Notice that Proposition 2.1 provides a guideline for the asymptotic error. In practice, although we don't know the value of $\tau_\infty^{\mathbf{w}}$, we can always maximize the finite sample SNR $\tau_n^{\mathbf{w}}$ to achieve asymptotic optimality, which has an explicit solution as shown in Theorem 2.2. The optimization of $\tau_n^{\mathbf{w}}$ is just calculation and we will omit the proof. It's also worth noticing that when $\Omega^{\mathbf{w}}$ is fixed, $\tau_\infty^{\mathbf{w}} = \infty$, in which the asymptotic error will always be 1. As we will see in Section 2.4, optimizing $\tau_n^{\mathbf{w}}$ would always give us a better finite sample result no matter whether the limit $\tau_\infty^{\mathbf{w}}$ is the same or not.

Theorem 2.2. *For balanced $A^{[L]} \sim \mathrm{MPPM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$ under Condition 2.1 and 2.2, the optimal weight $\mathbf{w}^* \in \mathcal{W}_L$ that minimizes the asymptotic spectral clustering error satisfies*

$$w_l^* \propto \frac{\Omega_{in}^{(l)} - \Omega_{out}^{(l)}}{\Omega_{in}^{(l)} + (K-1)\Omega_{out}^{(l)}}, \text{ for } l \in [L]. \tag{2.1}$$

Theorem 2.2 is illuminating because of its simple and intuitive form. We can see that the optimal weight for each layer is only determined by the layer's parameters. The right-hand side of (2.1) can be considered as the SNR of each single-layer network. The larger the difference between the within- and between-community probability is compared with the average connection probability, the larger the weight should be. This indicates that to make full use of the information, we should put more weights on the more informative layer. For balanced MPPM, Theorem 2.2 provides a closed-form solution of the optimal weights, so we could iteratively estimate $\hat{\mathbf{w}}$ and $\hat{\Omega}^{[L]}$.

Algorithm 2.1 Closed-form solution optimization

```
Input: adjacency matrices $A^{[L]}$, number of communities $K$ and precision parameter $\delta$
Output: multilayer network weight $\hat{\mathbf{w}}$ and community estimation $\hat{\mathbf{c}}$
Apply spectral clustering on each single adjacency matrix $A^{(l)}$, compute connectivity matrix estimate $\hat{\Omega}^{(l)}$ and the corresponding weight estimate $\hat{w}_l$.
while $\|\hat{\mathbf{w}}_{\text{old}} - \hat{\mathbf{w}}_{\text{new}}\| > \delta$ do
    Apply spectral clustering on WAM $A^{\hat{\mathbf{w}}_{\text{old}}}$ and obtain the community estimate $\hat{\mathbf{c}}$.
    Compute the corresponding $\hat{\Omega}^{(l)}$ and weights $\hat{\mathbf{w}}_{\text{new}}$ based on $\hat{\mathbf{c}}$.
end while
```

However, as we will see in Section 2.4, Algorithm 2.1 may easily fail when the true model violates the MPPM setting. To handle more complicated cases, some intrinsic properties should be explored, as it is discussed in the next subsection.

2.3.2 Eigenvalue gap optimization

We take a detour by looking at the spectrum of $A^{\mathbf{w}}$, then we will see the monotonic relation between SNR $\tau_n^{\mathbf{w}}$ and the eigenvalue gap $\lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}}$, where $\lambda_i^{\mathbf{w}}$ is ith largest eigenvalue of $A^{\mathbf{w}}$ in magnitude. By combining the semicircle law [54] and matrix perturbation theory [23], we could prove the following proposition.

Proposition 2.3. *For balanced $A^{[L]} \sim \mathrm{MPPM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$ under Condition 2.1 and 2.2, if $\mathbf{w} \in \mathcal{W}_L$, we have*

$$\lim_{n\to\infty} \frac{\lambda_K^{\mathbf{w}}}{\lambda_{K+1}^{\mathbf{w}}} = \begin{cases} \tau_\infty^{\mathbf{w}} + (4\tau_\infty^{\mathbf{w}})^{-1}, & \text{if } \tau_\infty^{\mathbf{w}} > 1/2, \\ 1, & \text{if } \tau_\infty^{\mathbf{w}} \le 1/2. \end{cases}$$

Specifically, when $\tau_\infty^{\mathbf{w}} = \infty$, $\lim_{n\to\infty} \lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}} = \infty$.

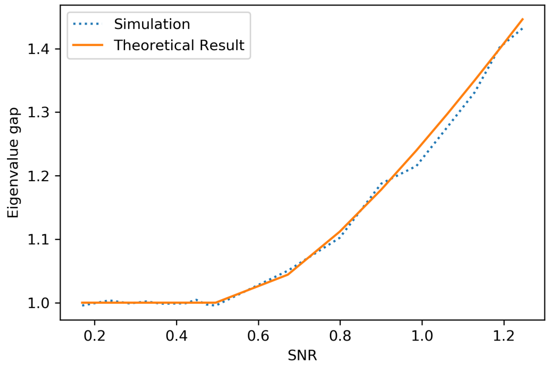


Figure 2.1: Eigenvalue gap versus SNR

Figure 2.1 shows the simulation and theoretical results from a PPM, in which $1/2$ is obviously the critical point. Proposition 2.3 tells us that when $\tau_\infty^{\mathbf{w}} > 1/2$, we could maximize the eigenvalue gap $\lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}}$ to maximize $\tau_n^{\mathbf{w}}$, which leads to the following theorem.

Theorem 2.4. *For balanced $A^{[L]} \sim \mathrm{MPPM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$ under Condition 2.1 and 2.2, if*

$$\lim_{n\to\infty} \frac{n\rho_n}{K} \sum_{l=1}^{L} \frac{\Omega_{in}^{(l)} - \Omega_{out}^{(l)}}{\Omega_{in}^{(l)} + (K-1)\Omega_{out}^{(l)}} > 1,$$

minimizing the asymptotic clustering error is equivalent to maximizing the eigenvalue gap $\lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}}$ for $\mathbf{w} \in \mathcal{W}_L$.

The additional condition in Theorem 2.4 is obtained by taking the limit of $\tau_n^{\mathbf{w}^*}$. Only requiring the maximum of $\tau_\infty^{\mathbf{w}}$ to be larger than $1/2$ is enough for us to apply Proposition 2.3. By Theorem 2.4, we define the objective function $g(\mathbf{w}) \equiv \lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}}$, and maximize it using Algorithm 2.2.

We use the explicit formula provided in [99] to compute the gradient of eigenvalues, which is $d\lambda_i = \mathbf{u}_i^T dA\mathbf{u}_i$ in our case. Taking the constraint $\|\mathbf{w}\|_1 = 1$ into consideration and applying the chain rule, we can derive $d\lambda_i^{\mathbf{w}}/dw_l = \mathbf{u}_i^T\{A^{(l)} - (L-1)^{-1}\sum_{l'\neq l} A^{(l')}\}\mathbf{u}_i$, so the gradient of $g(\mathbf{w})$ is

$$\nabla g(\mathbf{w}) = (\lambda_{K+1}^{\mathbf{w}}\nabla\lambda_K^{\mathbf{w}} - \lambda_K^{\mathbf{w}}\nabla\lambda_{K+1}^{\mathbf{w}})/{\lambda_{K+1}^{\mathbf{w}}}^2.$$

Algorithm 2.2 Eigenvalue gap optimization

```
Input: adjacency matrices $A^{[L]}$, number of communities $K$, initial learning rate $\gamma_0$, decay rate $r$,
    maximum iteration $T$ and weight precision parameter $\delta$
Output: multilayer network weight $\hat{\mathbf{w}}$ and community estimation $\hat{\mathbf{c}}$
Apply spectral clustering on each single $A^{(l)}$, compute connectivity matrix estimate $\hat{\Omega}^{(l)}$ and
    the corresponding weight estimate $\hat{w}_l$.
while $\|\hat{\mathbf{w}}_{\text{old}} - \hat{\mathbf{w}}_{\text{new}}\| \leq \delta$ and $t \leq T$ do
    Calculate the eigendecomposition of $A^{\hat{\mathbf{w}}_{\text{old}}}$.
    Use the gradient descent method on $\lambda_K^{\hat{\mathbf{w}}}/\lambda_{K+1}^{\hat{\mathbf{w}}}$ to update $\hat{\mathbf{w}}_{\text{new}}$.
end while
```

By gradient descent algorithm, we update $\mathbf{w}$ using $\mathbf{w}_{t+1} = \mathbf{w}_t + \gamma_t \nabla g$, where $\gamma_t = \gamma_0/(1 + rt)$ is the learning rate decaying with iterations, r is the dacay rate and t is the number of iterations. To avoid the algorithm converges to local maximum, we also need different random initial values for $\mathbf{w}$. It's worth noticing that although in this paper we only prove Theorem 2.4 for balanced MPPM, it's very robust under misspecified models as we will see in Section 2.4.2.

2.4 Simulations

We compare our methods with three different methods, which are mean adjacency matrix (Mean adj.) [69], aggregate spectral kernel (SpecK) [114] and module allegiance matrix (Module alleg.) [32]. We use 'Algo_fm' and 'Algo_eg' to refer Algorithm 2.1 and 2.2 respectively. For underlying models, we consider MPPM and MSBM to check the performance of our algorithms under different settings.

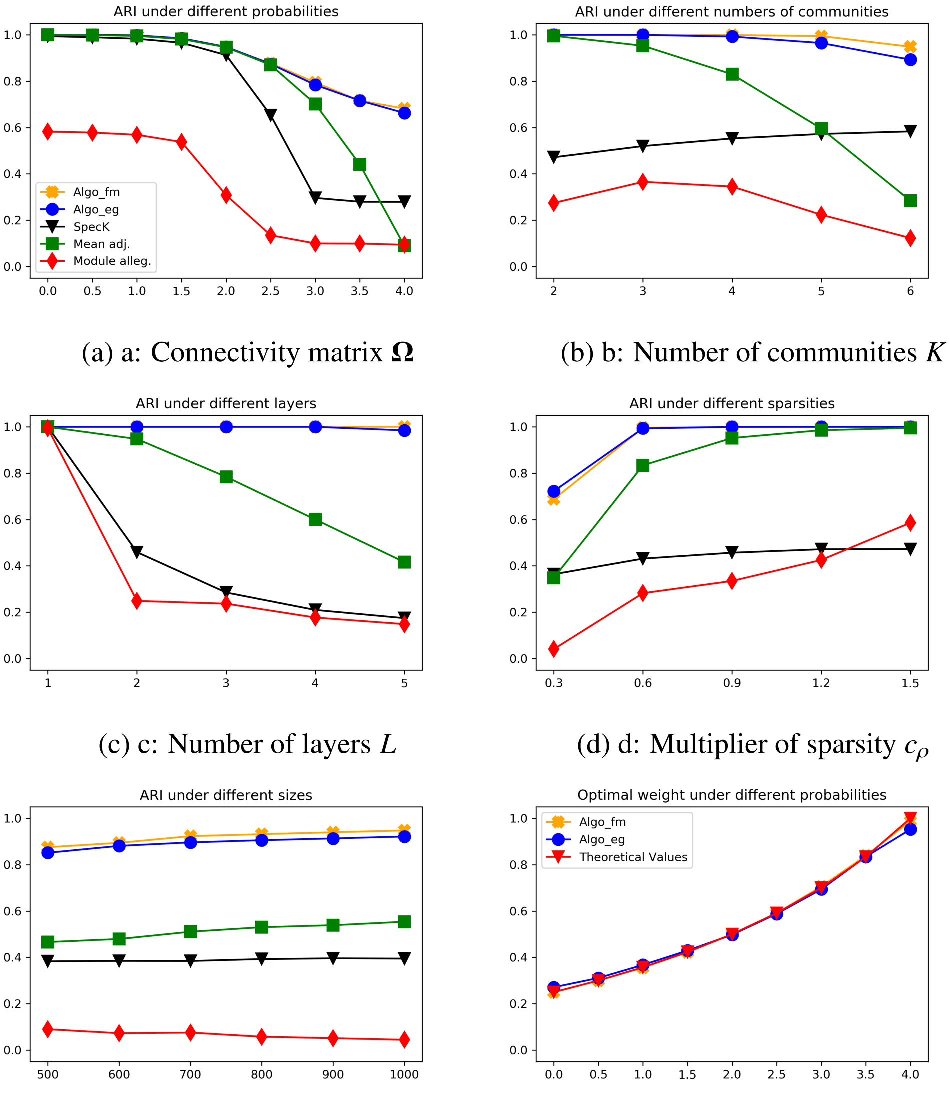
2.4.1 MPPM setting

For balanced MPPM($\Omega^{[L]}, \mathbf{c}, \rho_n$), there are at least five parameters to tune, which are Ω, K, L, n, ρ_n. We take $\rho_n = c_\rho \log n/n$, and only change c_ρ in the experiments when tuning ρ_n.

For Ω_{out} column in the above table, we list two parameters respectively for two layers in experiment a,b,d,e. In experiment c, 0 is used for 1 layer, all other layers use 4. We can see that our algorithms perform best under different settings from Figure 2.2, especially when there are more noise layers. Under the balanced MPPM, Algorithm 2.1 and 2.2 perform alike.

Table 2.1: Parameters of experiments under MPPM

Experiment	n	K	L	c_ρ	Ω_{in}	Ω_{out}
a	600	2	2	1.5	4	2/0-4
b	600	2-6	2	1.5	4	0/3
c	600	2	1-5	1.5	4	0/4
d	600	2	2	0.3-1.5	4	0/3
e	500-1000	2	2	0.3	4	0/3



(a) a: Connectivity matrix $\boldsymbol{\Omega}$

(b) b: Number of communities K

(c) c: Number of layers L

(d) d: Multiplier of sparsity c_ρ

(e) e: Number of nodes n

(f) Weight in Experiment a

Figure 2.2: ARI results from MPPM experiments

In Experiment a, we change the between-group probability of a single layer. A larger Ω_{out} introduces more noises. As we can imagine, with the number of layers fixed, ARI decreases with the increase of noises for all algorithms. However, compared with Mean adj., Algorithm 2.1 and 2.2 are both very robust against the increase of noises in one single layer since they will adaptively downweigh the noise layers. Figure 2.2f shows the optimal weight for the first layer learned from our algorithms matches the theoretical oracle value. When we add the number of communities in Experiment b, Mean adj. becomes worse. The reason is that when community numbers increase with everything else fixed, the decay of effective information in each layer is different, so the optimal weights change as well. This can be seen from Equation (2.1). A layer with lower between-group noises should be weighted more when the number of communities increases. As we fix one layer to be informative while adding more noise layers in Experiment c, our algorithms are very robust even when multiple noise layers exist. As we can imagine, the Mean adj. cannot handle noise layers properly. Although our algorithms and Mean adj. are all consistent in dense cases, our algorithms still outperform Mead adj. in sparse cases in Experiment d. With the increase of n in Experiment e, although Mean adj. is also consistent, our algorithms are better in finite samples.

Instead of fixing $\Omega^{[L]}$, we also consider the following three cases corresponding to Proposition 2.1. Algorithm 2.2 and Mean adj. are compared.

- High ($\tau_\infty^{\mathbf{w}} = \infty$): $\Delta_n = (\log 100)^{-1/2}$
- Medium ($0 < \tau_\infty^{\mathbf{w}} < \infty$): $\Delta_n = (\log n)^{-1/2}$
- Low ($\tau_\infty^{\mathbf{w}} = 0$): $\Delta_n = \{\log(0.01 n^2)\}^{-1/2}$

Table 2.2: Parameters of experiments under MPPM for different scales

Experiment	n	K	L	c_ρ	Ω
f	100-4600	2	2	2	$2J_2 + \begin{pmatrix} 2\Delta_n & -2\Delta_n \\ -2\Delta_n & 2\Delta_n \end{pmatrix}, 2J_2 + \begin{pmatrix} \Delta_n/2 & -\Delta_n/2 \\ -\Delta_n/2 & \Delta_n/2 \end{pmatrix}$

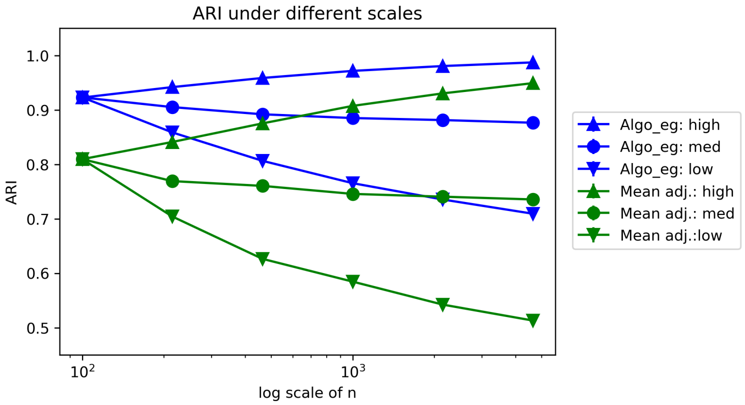


Figure 2.3: ARI results under different scales

Figure 2.3 validates what we claim in Proposition 2.1. In all cases, Algorithm 2.2 is better than Mean adj. We can see that in the high SNR case, the error rates for both algorithms converge to 0.5. In the low SNR case, the error rates increase as n grows, but Algorithm 2.2 is much more robust. For the medium SNR case, the accuracy of both algorithms converge to some constant, while Algorithm 2.2 is asymptotically better.

2.4.2 MSBM setting

Besides the MPPM setting, we also apply our algorithms under balanced MSBM to test the robustness under misspecified models. We have three MSBM models as described in Table 2.3. In Experiment g1 and g2, we have two layers MSBM, each with two communities. One layer has parameters that deviate a lot from PPM, which could potentially deteriorate the performance of Algorithm 2.1. As we can see in Figure 2.4a, Algorithm 2.1 always underperforms Algorithm 2.2 and Mean adj. except when the network is dense, so even suboptimal weights would yield good result. In this situation, using a simple mean weight is good enough.

However, applying a simple mean weight won't give us optimal results when there are more noise layers, just like Experiment h1, h2, i1 and i2. As we can see, in all those four experiments, even though the informative layers are against PPM settings, Algorithm 2.2 still consistently outperforms Mean adj. The performance of Algorithm 2.1 is very interesting. While it's similar or even slightly better than Algorithm 2.2 in Experiment i1 and i2, it's not consistent in Experiment

Table 2.3: Parameters of experiments under MSBM

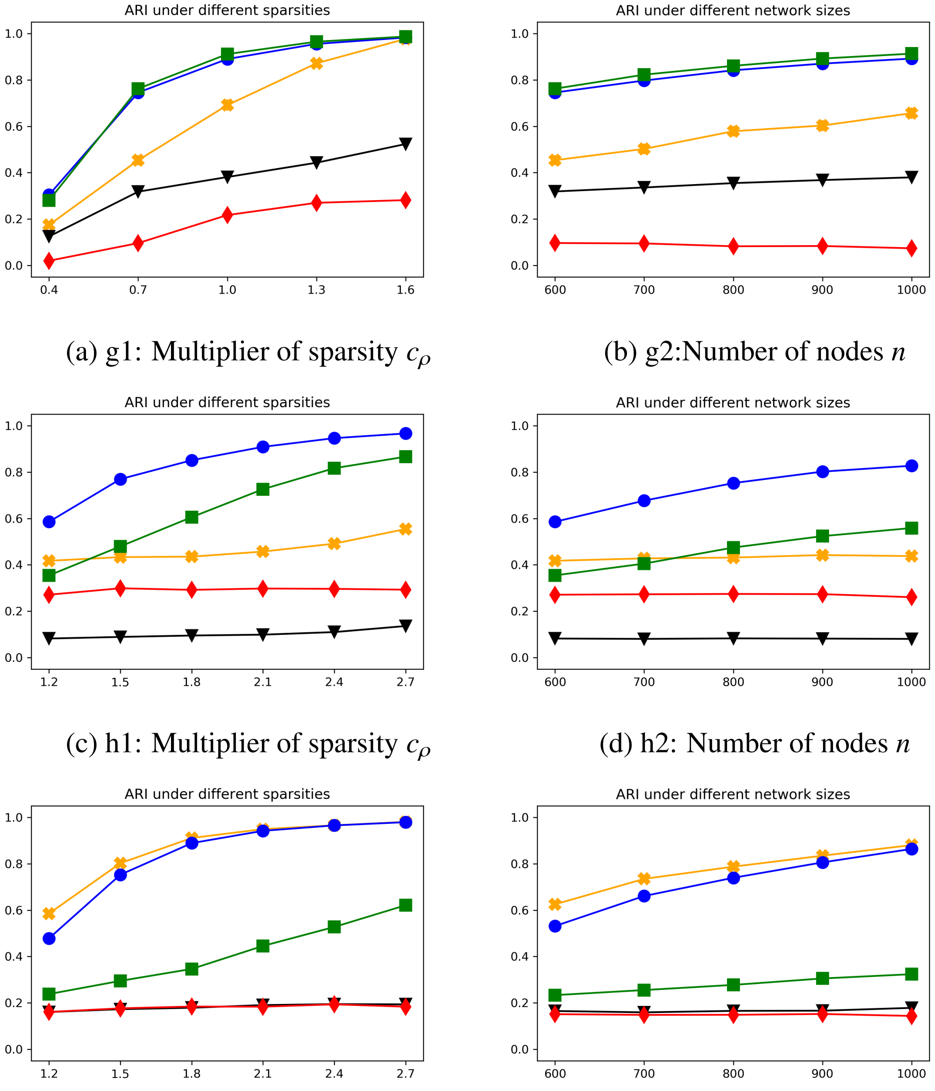
Experiment	n	K	L	c_ρ	Ω
g1	600	2	2	0.4-1.6	$\begin{pmatrix}10 & 2\\ 2 & 3\end{pmatrix},\begin{pmatrix}3 & 2\\ 2 & 3\end{pmatrix}$
g2	600-1000			0.7	
h1	600	3	4	1.2-2.7	$\begin{pmatrix}9 & 2 & 2\\ 2 & 2 & 2\\ 2 & 2 & 9\end{pmatrix},\begin{pmatrix}2 & 2 & 2\\ 2 & 4 & 2\\ 2 & 2 & 2\end{pmatrix},2J_3,2J_3$
h2	600-1000			1.2	
i1	600	3	4	1.2-2.7	$\begin{pmatrix}8 & 2 & 2\\ 2 & 6 & 2\\ 2 & 2 & 4\end{pmatrix},2J_3,2J_3,2J_3$
i2	600-1000			1.2	

h1 and h2. The reason is that while Algorithm 2.2 is pretty robust when more noise layers present, it can't handle the case when SBM is too far away from PPM setting. This is straightforward to see because Algorithm 2.1 relies heavily on PPM assumptions to derive the optimal weights. Although we only prove the consistency of Algorithm 2.2 under MPPM, the simulation results indicate that the theory may also hold for more general cases.

2.5 Real data example

We apply our algorithms to S&P 1500 data to see whether we could get extra information from combining multilayer networks in practice. S&P 1500 index should be a good representative of the US economy since it covers 90% of the market value of US stocks. We get the daily adjusted close price of stocks from Yahoo! Finance, of which dates range from 2001-01-01 to 2019-06-30, including 4663 trading days in total. We keep only stocks with less than 50 days' missing data and forward fill the price, which leaves us with 1020 stocks. According to the newest Global Industry Classification Standard, there are in total of 11 sectors, which we treat as the hidden ground truth that we hope to discover from stock prices. We remove sector 'Communication Services' because of small sizes and sectors 'Industrials' and 'Materials' because of their similar performances during economic cycles. The final dataset contains 770 stocks from 8 sectors.

We use the Pearson correlation of log returns between stocks to construct the network. The



(a) g1: Multiplier of sparsity c_ρ

(b) g2:Number of nodes n

(c) h1: Multiplier of sparsity c_ρ

(d) h2: Number of nodes n

(e) i1: Multiplier of sparsity c_ρ

(f) i2: Number of nodes n

Figure 2.4: ARI results from MSBM experiments

usage of log returns is the convention for stock prices. The correlation usage is based on the belief that stocks from the same sector should yield more similar returns because of their intrinsic similarities. Although the correlation is a continuous random variable ranging from -1 to 1, which contradicts the MSBM setting, it's the most natural way to construct the network. Since different sectors have a different baseline of correlations, we'll lose a lot of information if applying a uniform threshold to transform the correlation matrix into a matrix with only 0 and 1 entries. We could use this example to test the robustness of our algorithms to non-Bernoulli distributions. The most

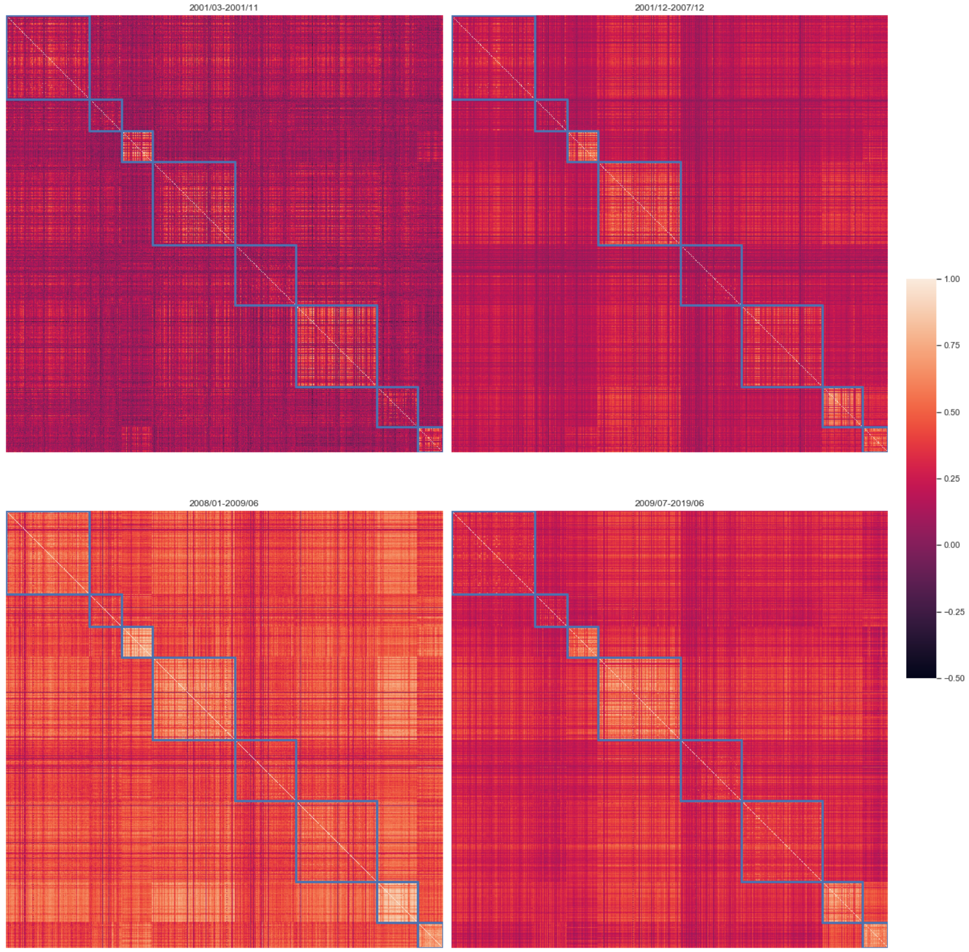
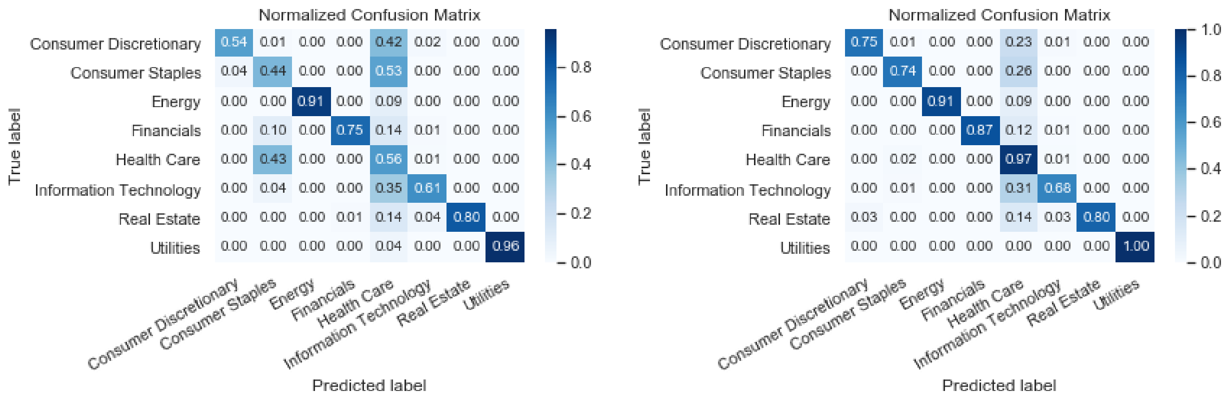


Figure 2.5: Correlation adjacency matrices for different time windows

straightforward idea to build the adjacency matrix is to use the whole time window to calculate the correlation. However, since financial data is usually non-stationary, correlation is very likely to change through time. In that case, connection probability may change over time, which is even against the assumption of general SBM.

With this intuition, we split the data into four time periods according to the National Bureau of Economic Research, which are respectively recession I (2001/03-2001/11), expansion I (2001/12-2007/12), recession II (2008/01-2009/06) and expansion II (2009/07-2019/06). The intuition is that the economy cycle is a determinant of sector performance over the intermediate-term. Different sectors tend to perform differently compared with the market in different phases of the economy. For example, ‘consumer discretionary’ tends to outperform the market in the early cycle. Also, even within the same sector, correlation can vary a lot at different times, which could be handled by our model. Taking a look at the correlation adjacency matrix in Figure 2.5, in which the box represents different sectors. We could observe that the correlations vary during different time windows. In the second recession period, we could observe an obviously higher correlation between or within all sectors, which is exactly what we expect. Also, for the expansion period, there are larger differences both within and between sectors. This demonstrates the necessity to use a multilayer model in this example.



True label \ Predicted label	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Information Technology	Real Estate	Utilities
Consumer Discretionary	0.54	0.01	0.00	0.00	0.42	0.02	0.00	0.00
Consumer Staples	0.04	0.44	0.00	0.00	0.53	0.00	0.00	0.00
Energy	0.00	0.00	0.91	0.00	0.09	0.00	0.00	0.00
Financials	0.00	0.10	0.00	0.75	0.14	0.01	0.00	0.00
Health Care	0.00	0.43	0.00	0.00	0.56	0.01	0.00	0.00
Information Technology	0.00	0.04	0.00	0.00	0.35	0.61	0.00	0.00
Real Estate	0.00	0.00	0.00	0.01	0.14	0.04	0.80	0.00
Utilities	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.96

(a) Single-layer network

True label \ Predicted label	Consumer Discretionary	Consumer Staples	Energy	Financials	Health Care	Information Technology	Real Estate	Utilities
Consumer Discretionary	0.75	0.01	0.00	0.00	0.23	0.01	0.00	0.00
Consumer Staples	0.00	0.74	0.00	0.00	0.26	0.00	0.00	0.00
Energy	0.00	0.00	0.91	0.00	0.09	0.00	0.00	0.00
Financials	0.00	0.00	0.00	0.87	0.12	0.01	0.00	0.00
Health Care	0.00	0.02	0.00	0.00	0.97	0.01	0.00	0.00
Information Technology	0.00	0.01	0.00	0.00	0.31	0.68	0.00	0.00
Real Estate	0.03	0.00	0.00	0.00	0.14	0.03	0.80	0.00
Utilities	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b) Multilayer network by Algorithm 2.2

Figure 2.6: Confusion matrices of different clustering methods

The ARIs from Algorithm 2.1, Algorithm 2.2, Mean adj., SpecK, Module alleg. are respectively 0.35, 0.58, 0.37, 0.44 and 0.33, while the ARI from spectral clustering on the whole time window is 0.41. We can see only Algorithm 2.2 significantly outperforms the baseline, which shows its robustness. One thing worth noticing is that the weights for four layers are and [0.56,0.21,0.07,0.16] by Algorithm 2.1 and [0.075,0.324,0.000,0.601] by Algorithm 2.2, which shows the relative importance in clustering for different time windows. The weight learned from Algorithm 2.2 tells us expansions play a much more important role compared with recessions. This is in accordance with our knowledge that in recession different sectors collapse in a similar way, but they will grow in different paces during expansion. Figure 2.6 shows the confusion matrix of using single-layer network versus multilayer network by Algorithm 2.2, from which we can observe significant accuracy improvement for consumer discretionary, consumer staples and health care sectors by using multilayer networks.

2.6 Discussion

Although we only prove Theorem 2.4 under balanced MPPM setting, we see the potential of Algorithm 2.2 under more general models from both simulations and real data. For future researches, one possibility is to extend the theoretical results to general SBM. Also, we only consider assortative networks here. For disassortative cases, we may introduce negative weights.

We see a phase transition phenomenon in Proposition 2.3, and there are two possible research directions that originated from here. One is whether below the threshold the problem is unsolvable by spectral clustering. The other is whether the threshold we derive here is a threshold for all algorithms or just for spectral clustering.

2.7 Proofs

The proofs are organized according to the order of the results in Section 2.3.

2.7.1 Proof of Proposition 2.1

To derive the error rate, we need to study the asymptotic distribution of eigenvectors first. Similar to [129], we define the spectral embedding for general SBM. For a fixed positive integer $d \leq n$, we call the $n \times d$ matrix $\hat{X} = U_A \Lambda_A^{1/2}$ the adjacency spectral embedding (ASE) of A into $\mathbb{R}^d$. We call $X = \rho_n^{-1/2} U_P \Lambda_P^{1/2}$ and $\nu = U_\Omega \Lambda_\Omega^{1/2}$ respectively the probability spectral embedding (PSE) and connectivity spectral embedding (CSE). Let $\mu^{\mathbf{w}}$ denote the CSE of $\Omega^{\mathbf{w}}$, then

$$\mu^{\mathbf{w}} \mu^{\mathbf{w}T} = \Omega^{\mathbf{w}} = \sum_{l=1}^{L} w_l \Omega^{(l)} = \sum_{l=1}^{L} w_l \nu^{(l)} \nu^{(l)^T}.$$

The entry of $\Omega^{\mathbf{w}}$ is $\Omega^{\mathbf{w}}_{ij} = \sum_{l=1}^{L} w_l \nu_{i\cdot}^{(l)^T} \nu_{j\cdot}^{(l)} = \mu_{i\cdot}^{\mathbf{w}T} \mu_{j\cdot}^{\mathbf{w}}$. Similarly, define $P^{\mathbf{w}} = X^{\mathbf{w}} X^{\mathbf{w}T}$.

Remark 2.1. *When Ω is positive semi-definite, which is true for assortative SBM since it's a Gram matrix. By taking $d = K$, we have $\Omega = \nu\nu^T$. Since P is an augmentation of Ω, we know $\Lambda_P = \Lambda_\Omega$*

and rows of U_P are just replications of rows of U_Ω. We will still keep d instead of K in later statements to avoid the confusion of columns and rows.

Remark 2.2. *The PPM assumes homogeneous probability across different communities. In this case, we know the exact eigenvalues are $\lambda_1(\Omega) = \Omega_{in} + (K-1)\Omega_{out}$ and $\lambda_2(\Omega) = \cdots = \lambda_d(\Omega) = \Omega_{in} - \Omega_{out}$. When $\Omega_{in} > \Omega_{out}$, Ω is full rank. The orthogonal eigenvectors are chosen as*

$$U_\Omega = \begin{bmatrix} K^{-1/2} & -\{(K-1)/K\}^{1/2} & 0 & \cdots & 0 \\ K^{-1/2} & \{K(K-1)\}^{-1/2} & -\{(K-2)/(K-1)\}^{1/2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ K^{-1/2} & \{K(K-1)\}^{-1/2} & \{(K-1)(K-2)\}^{-1/2} & \cdots & 2^{-1/2} \end{bmatrix} \in \mathbb{R}^{K\times d},$$

then

$$\nu = \begin{bmatrix} (\lambda_1/K)^{1/2} & -\{\lambda_2(K-1)/K\}^{1/2} & 0 & \cdots & 0 \\ (\lambda_1/K)^{1/2} & [\lambda_2/\{K(K-1)\}]^{1/2} & -\{\lambda_3(K-2)/(K-1)\}^{1/2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (\lambda_1/K)^{1/2} & [\lambda_2/\{K(K-1)\}]^{1/2} & [\lambda_3/\{(K-1)(K-2)\}]^{1/2} & \cdots & (\lambda_d/2)^{1/2} \end{bmatrix} \in \mathbb{R}^{K\times d}.$$

In [129], the eigenvector distribution for SBM has been discussed thoroughly, the question remains to be whether the eigenvector distribution of WAM still asymptotically follows a multivariate normal distribution. This result will be validated and expanded here, which is organized as follows. First, we introduce a direct consequence of Corollary 2.3 in [129] and prove a parallel theorem for WAM. Then, we'll prove a theorem when the connection of the random graph changes with n rather than fixed as usual. We will only focus on the case when $\rho_n \to 0$ because $\rho_n = 1$ is trivial in terms of the clustering error rate.

Theorem 2.5 (ASE convergence)**.** *Let $A^{[L]} \sim \mathrm{MSBM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$. For any fixed $\mathbf{w} \in \mathcal{W}_L$, consider the spectral embedding of $A^{\mathbf{w}}$. There exists an orthogonal matrix $W_n \in \mathbb{R}^{d\times d}$ and a matrix $R_n \in n \times d$ such that*

$$\hat{X}^{\mathbf{w}} W_n - \rho_n^{1/2} X^{\mathbf{w}} = \rho_n^{-1/2}(A^{\mathbf{w}} - P^{\mathbf{w}}) X^{\mathbf{w}} (X^{\mathbf{w}T} X^{\mathbf{w}})^{-1} + R_n.$$

Also, $\|R_n\| = O((n\rho_n)^{-1/2})$*. Let* $\boldsymbol{\theta}_l = \mathbb{E}[X_{1\cdot}^{(l)}]$ *for* $l \in [L]$ *and* $\Delta_{\mathbf{w}} = \mathbb{E}[X_{1\cdot}^{\mathbf{w}} X_{1\cdot}^{\mathbf{w}T}]$*. If Condition 2.1 holds, then there exists a sequence of orthogonal matrices* $W_n \in \mathbb{R}^{d\times d}$ *satisfying*

$$\|\hat{X}^{\mathbf{w}} W_n - \rho_n^{1/2} X^{\mathbf{w}}\|_F^2 \xrightarrow{a.s.} \text{trace}\left[\Delta_{\mathbf{w}}^{-1}\mathbb{E}\left\{X_{1\cdot}^{\mathbf{w}} X_{1\cdot}^{\mathbf{w}T}\left(\sum_{l=1}^{L} w_l^2 X_{1\cdot}^{(l)^T}\boldsymbol{\theta}_l\right)\right\}\Delta_{\mathbf{w}}^{-1}\right].$$

Theorem 2.6 (WAM eigenvector distribution)**.** *Let* $A^{[L]} \sim \mathrm{MSBM}(\Omega^{[L]}, \mathbf{c}, \rho_n)$*, then for any fixed* $\mathbf{w} \in \mathcal{W}_L$*, consider the spectral embedding of* $A^{\mathbf{w}}$*. If Condition 2.1 holds, then for any fixed index* i*, given* $c_i = k$*, there exists a sequence of orthogonal matrices* W_n *satisfying*

$$n^{1/2}(W_n \hat{X}_{i\cdot}^{\mathbf{w}} - \rho_n^{1/2} X_{i\cdot}^{\mathbf{w}}) \xrightarrow{d} N(0, \Sigma_k),$$

where $\Sigma_k = \Delta_{\mathbf{w}}^{-1}\mathbb{E}[X_{1\cdot}^{\mathbf{w}} X_{1\cdot}^{\mathbf{w}T} \sum_{l=1}^{L} w_l^2 v_{k\cdot}^{(l)^T} X_{1\cdot}^{(l)}]\Delta_{\mathbf{w}}^{-1}$*.*

Corollary 2.7 (SBM eigenvector distribution)**.** *Let* $A \sim \mathrm{SBM}(\Omega, \mathbf{c}, \rho_n)$*, if Condition 2.1 holds, then there exists a sequence of orthogonal matrices* W_n *such that for any fixed index* i *that* $X_{i\cdot} = v_{k\cdot}$*,*

$$n^{1/2}(W_n \hat{X}_{i\cdot} - \rho_n^{1/2} X_{i\cdot}) \xrightarrow{d} N(0, \Sigma_k),$$

where $\Sigma_k = \Delta^{-1}\mathbb{E}[(v_{k\cdot}^T X_{1\cdot}) X_{1\cdot} X_{1\cdot}^T]\Delta^{-1} = \Delta^{-1}(\sum_{i=1}^{K} \pi_i \Omega_{ik} v_{i\cdot} v_{i\cdot}^T)\Delta^{-1}$ *and* $\Delta = \mathbb{E}[X_{1\cdot} X_{1\cdot}^T] = \sum_{k=1}^{K} \pi_k v_{k\cdot} v_{k\cdot}^T$*.*

Corollary 2.7 is straightforward from Theorem 2.6 since a single layer SBM is just a special case of WAM. It is exactly the same as what we could derive from Corollary 2.3 in [129] since all assortative SBM can be expressed in the form of the random dot product graph. The only difference between WAM and SBM is the variance. Since WAM changes the variance rather than the mean of the edge weight, this consequence should be expected.

From Theorem 2.6, we can see as long as v_k are distinct, the asymptotic eigenvector distribution will be different, which makes the clustering problem relatively easy. To provide more insights, we consider the case when the mean connection probabilities of different communities converge to the same value, i.e., the connectivity matrix Ω is allowed to change with n but with a finite limit Ω_∞, whose elements are all the same. By Remark 2.2, we see the first eigenvector is determined by the degree and doesn't contribute to community detection, so we consider the 'effective' CSE, which is the sub-matrix corresponding to $\lambda_2(\Omega^{\mathbf{w}})$ until $\lambda_d(\Omega^{\mathbf{w}})$. Since $\lambda_2(\Omega^{\mathbf{w}}) = \cdots = \lambda_d(\Omega^{\mathbf{w}})$, we could define the 'effective' CSE as

$$\begin{aligned}\tilde{\mu}^{\mathbf{w}} =& \sqrt{\lambda_2^{\mathbf{w}}}\begin{bmatrix} -\{(K-1)/K\}^{1/2} & 0 & \cdots & 0 \\ \{K(K-1)\}^{-1/2} & -\{(K-2)/(K-1)\}^{1/2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \{K(K-1)\}^{-1/2} & \{(K-1)(K-2)\}^{-1/2} & \cdots & 2^{-1/2} \end{bmatrix} \\ =& [\tilde{\boldsymbol{\mu}}_1, \cdots, \tilde{\boldsymbol{\mu}}_K]^T = \sqrt{\lambda_2^{\mathbf{w}}}[\mathbf{u}_1, \cdots, \mathbf{u}_K]^T \in \mathbb{R}^{K\times(d-1)},\end{aligned}$$

which is just a constant matrix multiplied by $\sqrt{\lambda_2^{\mathbf{w}}}$. We can see that for fixed K, $\mathbf{u}_1, \cdots, \mathbf{u}_K$ are just fixed constant vectors.

Notice that

$$\Delta_{\mathbf{w}} = \frac{1}{K}\sum_{k=1}^{K} \mu_{k\cdot}^{\mathbf{w}} \mu_{k\cdot}^{\mathbf{w}T} = \frac{1}{K}\mu^{\mathbf{w}T}\mu^{\mathbf{w}} = \frac{1}{K}D((\lambda_1(\Omega^{\mathbf{w}}), \cdots, \lambda_d(\Omega^{\mathbf{w}}))^T)$$

and

$$\nu_{k\cdot}^{(l)^T}\nu_{m\cdot}^{(l)} = \Omega_{km}^{(l)} = \begin{cases} \Omega_{in}^{(l)} \text{ if } k = m, \\ \Omega_{out}^{(l)} \text{ if } k \neq m. \end{cases}$$

So

$$\begin{aligned}&\mathbb{E}[X_{1\cdot}^{\mathbf{w}} X_{1\cdot}^{\mathbf{w}T} \sum_{l=1}^{L} w_l^2 \nu_{k\cdot}^{(l)^T} X_{1\cdot}^{(l)}] = \frac{1}{K}\sum_{m=1}^{K} \mu_{m\cdot}^{\mathbf{w}} \mu_{m\cdot}^{\mathbf{w}\,T} \sum_{l=1}^{L} w_l^2 \Omega_{km}^{(l)} \\ =& \frac{\Omega_{out}^{\mathbf{w}^2}}{K}\sum_{m=1}^{K} \mu_{m\cdot}^{\mathbf{w}} \mu_{m\cdot}^{\mathbf{w}\,T} + \frac{\Omega_{in}^{\mathbf{w}^2} - \Omega_{out}^{\mathbf{w}^2}}{K}\mu_{k\cdot}^{\mathbf{w}} \mu_{k\cdot}^{\mathbf{w}\,T} = \frac{\Omega_{out}^{\mathbf{w}^2}}{K}\mu^{\mathbf{w}T}\mu^{\mathbf{w}} + \frac{\Omega_{in}^{\mathbf{w}^2} - \Omega_{out}^{\mathbf{w}^2}}{K}\mu_{k\cdot}^{\mathbf{w}} \mu_{k\cdot}^{\mathbf{w}\,T} \\ =& \frac{\Omega_{out}^{\mathbf{w}^2}}{K}D((\lambda_1(\Omega^{\mathbf{w}}), \cdots, \lambda_d(\Omega^{\mathbf{w}}))^T) + \frac{\Omega_{in}^{\mathbf{w}^2} - \Omega_{out}^{\mathbf{w}^2}}{K}\mu_{k\cdot}^{\mathbf{w}} \mu_{k\cdot}^{\mathbf{w}\,T}.\end{aligned}$$

Then it's straight forward to see

$$\begin{aligned}\Sigma_k =&\Delta_{\mathbf{w}}^{-1}\left[\frac{\Omega_{out}^{\mathbf{w}^2}}{K}D((\lambda_1(\Omega^{\mathbf{w}}),\cdots,\lambda_d(\Omega^{\mathbf{w}}))^T)+\frac{\Omega_{in}^{\mathbf{w}^2}-\Omega_{out}^{\mathbf{w}^2}}{K}\mu_{k\cdot}^{\mathbf{w}}\mu_{k\cdot}^{\mathbf{w}\,T}\right]\Delta_{\mathbf{w}}^{-1}\\ =&K\Omega_{out}^{\mathbf{w}^2}D((1/\lambda_1(\Omega^{\mathbf{w}}),\cdots,1/\lambda_d(\Omega^{\mathbf{w}}))^T)+\frac{\Omega_{in}^{\mathbf{w}^2}-\Omega_{out}^{\mathbf{w}^2}}{K}\Delta_{\mathbf{w}}^{-1}\mu_{k\cdot}^{\mathbf{w}}\mu_{k\cdot}^{\mathbf{w}\,T}\Delta_{\mathbf{w}}^{-1}.\end{aligned}$$

We define $\tilde{\Sigma}_k$ to be the sub-matrix of Σ_k corresponding to $\tilde{\mu}^{\mathbf{w}}$. Also, we could define the limit 'effective' covariance matrix

$$\tilde{\Sigma} =\frac{\Omega_{in}^{\mathbf{w}^2}+(K-1)\Omega_{out}^{\mathbf{w}^2}}{\Omega_{in}^{\mathbf{w}}-\Omega_{out}^{\mathbf{w}}}I_{d-1}\equiv\sigma_{\mathbf{w}}^2 I_{d-1},$$

then it's easy to see that $\tilde{\Sigma}_k\tilde{\Sigma}^{-1}\xrightarrow{p} I$. Here we require that there $\exists\mathbf{w}$ s.t. $\sigma_{\mathbf{w}}\neq\infty$. This is equivalent to $\exists l\in[L]$ s.t. $\Omega_{in}^{(l)}\neq\Omega_{out}^{(l)}$, otherwise the problem is undetectable.

Based on Theorem 2.6, we could have the following theorem by the Slutsky's theorem.

Theorem 2.8 (WAM with varied mean)**.** *Let $A_n^{[L]}\sim\mathrm{MSBM}(\Omega_n^{[L]},\mathbf{c},\rho_n)$. For $l\in[L]$, assume $\lim_{n\to\infty}\Omega_n^{(l)}=\Omega_\infty^{(l)}\in\mathbb{R}^{K\times K}$, whose entries are all the same. Also, there $\exists l\in[L]$ s.t. $\Omega_{in}^{(l)}\neq\Omega_{out}^{(l)}$. For any fixed $\mathbf{w}\in\mathcal{W}_L$, if Condition 2.1 holds, there exists a sequence of orthogonal matrices W_n such that for any fixed index i that $c_i=k$,*

$$\sigma_{\mathbf{w}}^{-1}n^{1/2}(W_n\hat{X}_{i,2:d}^{\mathbf{w}}-\rho_n^{1/2}X_{i,2:d}^{\mathbf{w}})\xrightarrow{d}N(\mathbf{0},I_{d-1}).$$

Under the setting of Theorem 2.8, the asymptotic variance for different classes are all the same, which reduces the clustering problem from QDA to LDA. Also, the problem becomes non-trivial since the asymptotic means are all the same, so the error rate may not converge to 0 as before. We consider the probability of classifying a node to group k for $k\geq 2$ when it's from group 1 if only considering a two-class classification problem. The asymptotic distribution for $c_i=k$,

$$\frac{(n\rho_n\lambda_2^{\mathbf{w}})^{1/2}}{\sigma_{\mathbf{w}}}\{\hat{X}_{i,2:d}/(\rho_n\lambda_2^{\mathbf{w}})^{1/2}-\mathbf{u}_k\}\xrightarrow{d}N(\mathbf{0},I_{d-1}).$$

Since we use U_A for spectral clustering, here we focus on $\hat{X}_{i,2:d}/(\rho_n\lambda_2^{\mathbf{w}})^{1/2}$, whose density function under group k is

$$f_k(\mathbf{x}) = \frac{(n\rho_n\lambda_2^{\mathbf{w}})^{1/2}}{(2\pi)^{(d-1)/2}\sigma_{\mathbf{w}}} \exp\left\{-\frac{n\rho_n\lambda_2^{\mathbf{w}}}{2\sigma_{\mathbf{w}}^2}(\mathbf{x}-\mathbf{u}_k)^T(\mathbf{x}-\mathbf{u}_k)\right\}.$$

The probability of misclassifying k-th class to 1st class is $P(f_k \geq f_1|1) = \int_{\mathbf{x}\in\mathcal{R}_{k1}} f_1(\mathbf{x})d\mathbf{x}$, where $\mathcal{R}_{k1} = \left\{\mathbf{x}\in\mathbb{R}^{d-1} \mid (\mathbf{x}-\frac{1}{2}(\mathbf{u}_k+\mathbf{u}_1))^T(\mathbf{u}_k-\mathbf{u}_1)\geq 0\right\}$ is the region that the likelihood of group k is larger than group 1.

We can see that the region is a constant scale area and has nothing to do with $\Omega^{\mathbf{w}}$ and n. Since the centers $\mathbf{u}_1,\cdots,\mathbf{u}_k$ are all fixed, the error rate only depends on the SNR, which is $(n\rho_n\lambda_2^{\mathbf{w}})^{1/2}/\sigma_{\mathbf{w}} = 2K^{1/2}\tau_n^{\mathbf{w}}$. When K fixed, the error rate is monotonic decreasing with $\tau_n^{\mathbf{w}}$ and the conclusion in Proposition 2.1 can be seen. In general, we can always maximize $\tau_n^{\mathbf{w}}$ to minimize the asymptotic error. Specifically, when $\Omega^{\mathbf{w}}$ is fixed and $\Omega_{in}^{\mathbf{w}} \neq \Omega_{out}^{\mathbf{w}}$, the limiting error rate is always 0 since $\mathbf{u}_k$ are distinct, although we can still minimize finite sample error rate. When $\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}} = \Theta((n\rho_n)^{-1/2})$, which is case 3, we can minimize the error rate even in the asymptotic sense.

A. Proof of Theorem 2.5

Proof. We need to bound $\|A^{\mathbf{w}} - P^{\mathbf{w}}\|$ like Proposition 7 in [97]. We show a stronger result, which can be seen directly from Theorem 5.2 of [91]. Since we don't need to prove the theorem for the laplacian matrix, we only need ρ_n of order $\log n/n$ instead of $\log^4 n/n$.

Lemma 2.9. *For any constant $c, r > 0$, there exists constant $C(c,r) > 0$ independent of n and $\mathbf{w}$ such that the following holds. If $\|P^{\mathbf{w}}\|_1 > c\log n$, then with probability at least $1-n^{-r}$, the following hold*

$$\|A^{\mathbf{w}} - P^{\mathbf{w}}\| \leq C\sqrt{\|P^{\mathbf{w}}\|_1}.$$

Following the same proof procedure of Theorem A.5 in [128], we can prove the following lemma. Since all procedures are the same once Lemma 2.9 is established, we'll omit the proof of

Lemma 2.10.

Lemma 2.10. *Under the same assumption in Lemma 2.9, there exists an orthogonal matrix W_n such that for sufficiently large n,*

$$\|\hat{X}^{\mathbf{w}} - \rho_n^{1/2} X^{\mathbf{w}} W_n\| = \|(A^{\mathbf{w}} - P^{\mathbf{w}}) U_{P^{\mathbf{w}}} S_{P^{\mathbf{w}}}^{-1/2}\|_F + O\left(d\|P^{\mathbf{w}}\|_1^2 \lambda_d(P^{\mathbf{w}})^{-5/2}\right)$$

with high probability.

To verify Theorem 2.5, all steps are the same as Theorem 2.1 in [129], except for the calculation of $\mathbb{E}[(A^{\mathbf{w}} - P^{\mathbf{w}})^2]$, so we only show that part here. Since

$$\mathbb{E}\left[\sum_k (A_{ik}^{\mathbf{w}} - P_{ik}^{\mathbf{w}})(A_{kj}^{\mathbf{w}} - P_{kj}^{\mathbf{w}})\right] = \begin{cases} 0 & \text{if } i \neq j, \\ \sum_{k \neq i} \sum_{l=1}^{L} w_l^2 P_{kj}^{(l)} (1 - P_{kj}^{(l)}) & \text{if } i = j, \end{cases}$$

so

$$n^{-2} \rho_n^{-1} X^{\mathbf{w}T} \mathbb{E}\left[(A^{\mathbf{w}} - P^{\mathbf{w}})^2\right] X^{\mathbf{w}} = n^{-2} \rho_n^{-1} \sum_{i=1}^{n} X_{i\cdot}^{\mathbf{w}} X_{i\cdot}^{\mathbf{w}T} \sum_{k \neq i} \sum_{l=1}^{L} w_l^2 P_{ki}^{(l)} (1 - P_{ki}^{(l)})$$
$$= n^{-2} \sum_{i=1}^{n} X_{i\cdot}^{\mathbf{w}} X_{i\cdot}^{\mathbf{w}T} \sum_{k \neq i} \sum_{l=1}^{L} w_l^2 \{ X_{i\cdot}^{(l)^T} X_{k\cdot}^{(l)} - \rho_n (X_{i\cdot}^{(l)^T} X_{k\cdot}^{(l)})^2 \}$$

When $\rho_n \to 0$, the above term converges to

$$\mathbb{E}[X_{1\cdot}^{\mathbf{w}} X_{1\cdot}^{\mathbf{w}T} (\sum_l w_l^2 X_1^{(l)^T} \boldsymbol{\theta}_l)].$$

□

B. Proof of Theorem 2.6

Proof. The proof is similar to the proof of Theorem 2.2 in [129] except for the variance calculation. For any fixed index i we have

$$
\begin{aligned}
&\sqrt{n}(W_n\hat{X}^{\mathbf{w}}_{i\cdot} - \rho_n^{1/2}X^{\mathbf{w}}_{i\cdot}) = \sqrt{n}\rho_n^{-1/2}(X^{\mathbf{w}T}X^{\mathbf{w}})^{-1}\sum_{j\neq i}(A_{ij}-P_{ij})X^{\mathbf{w}}_{j\cdot} + o(1)\\
=&(n^{-1}X^{\mathbf{w}T}X^{\mathbf{w}})^{-1}\sum_{i\neq j}\frac{(A^{\mathbf{w}}_{ij}-\rho_n\sum_l w_l {X^{(l)}_{i\cdot}}^T X^{(l)}_{j\cdot})}{\sqrt{n\rho_n}}X^{\mathbf{w}}_{j\cdot}+o(1)
\end{aligned}
$$

Given $X^{\mathbf{w}}_{i\cdot} = \boldsymbol{\mu}_k$, $\sum_{i\neq j}(A^{\mathbf{w}}_{ij}-\rho_n\sum_l w_l {X^{(l)}_{i\cdot}}^T X^{(l)}_{j\cdot})X^{\mathbf{w}}_{j\cdot}$ is a sum of i.i.d. mean zero random variables. The conditional variance is calculated as

$$
\begin{aligned}
&\mathbb{E}[(A^{\mathbf{w}}_{ij}-\rho_n\sum_l w_l {\nu^{(l)}_{k\cdot}}^T X^{(l)}_{j\cdot})^2 X^{\mathbf{w}}_{j\cdot}X^{\mathbf{w}T}_{j\cdot}]\\
=&\mathbb{E}_m[\mathbb{E}[(A^{\mathbf{w}}_{ij}-\rho_n\sum_l w_l {\nu^{(l)}_{k\cdot}}^T X^{(l)}_{j\cdot})^2 X^{\mathbf{w}}_{j\cdot}X^{\mathbf{w}T}_{j\cdot}|X^{\mathbf{w}}_{j\cdot}=\boldsymbol{\mu}_m]]\\
=&\mathbb{E}_m[\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T\mathbb{E}[(A^{\mathbf{w}}_{ij}-\rho_n\sum_l w_l {\nu^{(l)}_{k\cdot}}^T X^{(l)}_{j\cdot})^2|X^{\mathbf{w}}_{j\cdot}=\boldsymbol{\mu}_m]]\\
=&\rho_n\mathbb{E}_m[\boldsymbol{\mu}_m\boldsymbol{\mu}_m^T\sum_{l=1}^{L} w_l^2 {\nu^{(l)}_{k\cdot}}^T\nu^{(l)}_{m\cdot}(1-\rho_n{\nu^{(l)}_{k\cdot}}^T\nu^{(l)}_{m\cdot})].
\end{aligned}
$$

This yields

$$
\sum_{i\neq j}\frac{(A^{\mathbf{w}}_{ij}-\rho_n\sum_l w_l {X^{(l)}_{i\cdot}}^T X^{(l)}_{j\cdot})}{\sqrt{n\rho_n}}X^{\mathbf{w}}_{j\cdot}\xrightarrow{d} N(0,\mathbb{E}[X^{\mathbf{w}}_{1\cdot}X^{\mathbf{w}T}_{1\cdot}\sum_{l=1}^{L} w_l^2{\nu^{(l)}_{k\cdot}}^T X^{(l)}_1(1-\rho_n{\nu^{(l)}_{k\cdot}}^T X^{(l)}_1)]).
$$

Since $n^{-1}X^{\mathbf{w}T}X^{\mathbf{w}}\xrightarrow{a.s.}\Delta_{\mathbf{w}}$ as $n\to\infty$, by Slutsky's theorem, when $\rho_n\to 0$,

$$
\sqrt{n}(W_n\hat{X}^{\mathbf{w}}_{i\cdot} - \rho_n^{1/2}X^{\mathbf{w}}_{i\cdot})\xrightarrow{d} N(0,\Delta_{\mathbf{w}}^{-1}\mathbb{E}[X^{\mathbf{w}}_{1\cdot}X^{\mathbf{w}T}_{1\cdot}\sum_{l=1}^{L} w_l^2{\nu^{(l)}_{k\cdot}}^T X^{(l)}_{1\cdot}]\Delta_{\mathbf{w}}^{-1}).
$$

□

2.7.2 Proof of Proposition 2.3

Results in this subsection will mostly be based on Proposition 2 in [15] with some slight modifications. By studying the asymptotic behavior of eigenvalue gap $\lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}}$, we will see clearly its connection with SNR $\tau_n^{\mathbf{w}}$, which results in Theorem 2.4. Since general SBM doesn't have an explicit solution for eigenvalues, we consider only balanced PPM here, whose eigenvalues have an explicit form.

We first make a decomposition as in [15] and state the spectrum property as follows.

Proposition 2.11 (WAM decomposition)**.** *Suppose $A^{\mathbf{w}}$ is the WAM generated from a balanced* MPPM. *Consider the scaled matrix $\hat{A}^{\mathbf{w}} = \gamma(n)A^{\mathbf{w}}$, where $\gamma(n) = K/\{n\rho_n(\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}})\}$. We decompose $\hat{A}^{\mathbf{w}}$ into two parts*

$$\hat{A}^{\mathbf{w}} = \tilde{A}^{\mathbf{w}} + \bar{A}^{\mathbf{w}},$$

where $\bar{A}^{\mathbf{w}} = \gamma(n)P^{\mathbf{w}}$ is the expectation of $\hat{A}^{\mathbf{w}}$ and $\tilde{A}^{\mathbf{w}}$ is a random matrix with zero mean random entries. The spectrum property of $\bar{A}^{\mathbf{w}}$ and $\tilde{A}^{\mathbf{w}}$ is summarized as follows.

1. *$\bar{A}^{\mathbf{w}}$ is of rank K and the 2nd to Kth largest eigenvalues all converge to 1.*
2. *The spectrum bound of $\tilde{A}^{\mathbf{w}}$ is $1/\tau_n^{\mathbf{w}}$, the reciprocal of SNR.*

According to [23], the scale of $\tau_n^{\mathbf{w}}$ determines the behavior of the 2nd to Kth eigenvalues, which is crucial for determining the spectrum of $\hat{A}^{\mathbf{w}}$. The property is summarized as follows.

Proposition 2.12 (WAM spectrum)**.** *Based on the decomposition in Proposition 2.11, $\tau_\infty^{\mathbf{w}}$ determines the asymptotic behavior of the spectrum of $\hat{A}^{\mathbf{w}}$ as follows.*

1. *If $\tau_\infty^{\mathbf{w}} = \infty$, the asymptotic spectrum of $\hat{A}^{\mathbf{w}}$ and $\bar{A}^{\mathbf{w}}$ are the same since the noises degenerate.*
2. *If $\tau_\infty^{\mathbf{w}} = 0$, the asymptotic spectrum of $\hat{A}^{\mathbf{w}}$ and $\tilde{A}^{\mathbf{w}}$ are the same since the noises dominate.*
3. *If $\tau_\infty^{\mathbf{w}}$ is a finite constant, the signals and noises are in the same scale. Specifically,*
   (a) *If $\tau_\infty^{\mathbf{w}} > 1/2$, the 2nd to Kth eigenvalue of $\hat{A}^{\mathbf{w}}$ converges to $1 + (4\tau_\infty^{\mathbf{w}})^{-2}$.*

(b) If $\tau_\infty^{\mathbf{w}} \leq 1/2$, the 2nd to Kth eigenvalue of $\hat{A}^{\mathbf{w}}$ converges to $\tau_\infty^{\mathbf{w}-1}$.

We can see that $\tau_\infty^{\mathbf{w}}$ is monotonic with the eigenvalue gap $\lambda_K^{\mathbf{w}}/\lambda_{K+1}^{\mathbf{w}}$. A larger $\tau_\infty^{\mathbf{w}}$ leads to a larger gap, except when $\tau_\infty^{\mathbf{w}} \leq 1/2$, in which case the noises still dominate signals. When $\tau_\infty^{\mathbf{w}} > 1/2$, the limit of eigenvalue gap converges to $(4\tau_\infty^{\mathbf{w}})^{-1} + \tau_\infty^{\mathbf{w}}$, which is monotonic increasing with $\tau_\infty^{\mathbf{w}}$. Thus, we can maximize the eigenvalue gap to maximize $\tau_n^{\mathbf{w}}$ asymptotically. This shows Theorem 2.4.

The proofs of Proposition 2.11 and 2.12 are mainly based on the semicircle law of $\tilde{A}^{\mathbf{w}}$ [15], and Theorem 2.1 in [23], which shows the spectrum property of the perturbations of large random matrices.

A. Proof of Proposition 2.11

Proof. By some algebra, it's easy to show $\bar{A}^{\mathbf{w}}$ is similar to some matrix with only $K \times K$ non-zero entries, whose eigenvalues are the same as the following $K \times K$ matrix

$$\begin{aligned} Y_n =& \frac{K}{n(\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}})} D((n_1, \cdots, n_K)^T)\Omega^{\mathbf{w}} \\ =& \frac{1}{\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}}}\Omega^{\mathbf{w}} + \frac{1}{\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}}} D((1 - Kn_1n^{-1}, \cdots, 1 - Kn_Kn^{-1})^T)\Omega^{\mathbf{w}}. \end{aligned}$$

By decomposing Y_n into two parts, we see that the 2nd to Kth eigenvalues of the first part is exactly 1. For the second part, since $1 - Kn_in^{-1} = \Theta(n^{-1/2})$ by CLT, as long as $\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}} = \omega(n^{-1/2})$, the second matrix converges to 0 elementwisely, so do all the eigenvalues. By Weyl's eigenvalue interlacing inequalities, the 2nd to Kth eigenvalues of Y_n converge to 1 as well.

Intuitively, the bound of $\tilde{A}^{\mathbf{w}}$ represents the magnitude of noises, which we hope to be as small as possible. To obtain the bound, we consider the following scaling and decomposition. Let $s_i^{\mathbf{w}} = \{\sum_{l=1}^L w_l^2 \sum_j P_{ij}^{(l)}(1 - P_{ij}^{(l)})\}^{1/2}$ be the empirical standard deviation of node i and $s^{\mathbf{w}} = \{n\rho_n(\Omega_{in}^{\mathbf{w}^2} + (K-1)\Omega_{out}^{\mathbf{w}^2})/K\}^{1/2}$ be the population standard deviation. Then

$$\frac{A^w - P^w}{s^{\mathbf{w}}} = \left[\frac{A_{ij}^{\mathbf{w}} - P_{ij}^{\mathbf{w}}}{s_i^{\mathbf{w}}}\right]_{n\times n} + \left[(A_{ij}^{\mathbf{w}} - P_{ij}^{\mathbf{w}})(\frac{1}{s_i^{\mathbf{w}}} - \frac{1}{s^{\mathbf{w}}})\right]_{n\times n}.$$

The first part is a generalized Wigner matrix, so its eigenvalues follow local semicircle law according to [54]. We claim that the spectrum bound of the second matrix converges to 0 asymptotically. Then by Weyl's inequality, we know the asymptotic spectrum distribution of $(A^w - P^w)/s^{\mathbf{w}}$ is not affected by the deformation, so it follows local semicircle law as well.

To validate the claim, we show a more general spectrum bound using the conclusion from [21]. Let $d_{\mathbf{w}} = \max_i \sum_{l=1}^{L} w_l^2 \sum_j P_{ij}^{(l)}$, $H^{\mathbf{w}} = d_{\mathbf{w}}^{-1/2}(A^{\mathbf{w}} - P^{\mathbf{w}})$, $q = [n\rho_n K^{-1}\{\Omega_{in}^{\mathbf{w}^2} + (K-1)\Omega_{out}^{\mathbf{w}^2}\}]^{1/2}$ and $\kappa = (\Omega_{in}^{\mathbf{w}^2}/\Omega_{out}^{\mathbf{w}})^2$. It's easy to verify that $d_{\mathbf{w}} \geq q^2$, $|H_{ij}^{\mathbf{w}}| \leq d_{\mathbf{w}}^{-1/2} \leq 1/q$ and $\mathbb{E}H_{ij}^{\mathbf{w}2} \leq \sum_{l=1}^{L} w_l^2 P_{ij}^{(l)}/d_{\mathbf{w}} \leq 1$, then $H^{\mathbf{w}}$ satisfies the following assumption.

Condition 2.3. *Let H be a symmetric random matrix whose upper triangular entries $(H_{ij})_{1\leq i\leq j\leq n}$ are independent mean-zero random variables. Moreover, suppose that there exist $q > 0$ and $\kappa \geq 1$ such that*

$$\max_i \sum_j \mathbb{E}[H_{ij}^2] \leq 1,\ \max_i \mathbb{E}[H_{ij}^2] \leq \frac{\kappa}{n},\ \max_{i,j} |H_{ij}| \leq \frac{1}{q}\ a.s.$$

Directly applying Theorem 2.6 in [21], we have the following theorem.

Theorem 2.13. *Assume Assumption 2.3 is satisfied by $H^{\mathbf{w}}$, then for $2 \vee q \leq n^{1/13}\kappa^{-1/12}$,*

$$\mathbb{E}\|H^{\mathbf{w}}\| \leq 2 + C\frac{\eta}{(1 \vee \log\eta)^{1/2}},\ \textit{with}\ \eta = \frac{(\log n)^{1/2}}{q},$$

for some universal constant $C > 0$. In particular,

$$\mathbb{E}\|H^{\mathbf{w}}\| \leq 2 + C\frac{(\log n)^{1/2}}{q}.$$

Theorem 2.13 shows that with high probability, $\|H^{\mathbf{w}}\| \leq 2 + o(1)$ as long as $q^2 \gg \log n$ and $\kappa \ll n^{12/13}$. By Theorem 2.2 of [81] and Corollary 1.4 of [22], we know this bound is actually sharp. Since $(s_i^{\mathbf{w}}s^{\mathbf{w}})^{-1}(s_i^{\mathbf{w}} - s^{\mathbf{w}})d_{\mathbf{w}}^{1/2} = \Theta(n^{-1/2})$, we know from Theorem 2.13 that the spectrum bound of the second matrix is $\Theta(n^{-1/2})$, which validates the claim.

From above discussion, we know immediately that $\|\tilde{A}^{\mathbf{w}}\|$ is bounded by

$$2s^{\mathbf{w}}\gamma(n) = 2\left(\frac{K}{n\rho_n}\right)^{1/2}\frac{\{\Omega_{in}^{\mathbf{w}^2} + (K-1)\Omega_{out}^{\mathbf{w}^2}\}^{1/2}}{\Omega_{in}^{\mathbf{w}} - \Omega_{out}^{\mathbf{w}}} = \frac{1}{\tau_n^{\mathbf{w}}}.$$

When $\tau_\infty^{\mathbf{w}}$ is finite, the semi-circle law holds for $\tilde{A}^{\mathbf{w}}$ with bound $1/\tau_\infty^{\mathbf{w}}$.

□

B. Proof of Proposition 2.12

Proof. Now we know $\tau_\infty^{\mathbf{w}}$ plays a critical role in determining the asymptotic spectrum distribution of $\hat{A}^{\mathbf{w}}$. When $\tau_\infty^{\mathbf{w}} = 0$, the 2nd to Kth eigenvalues of $\hat{A}^{\mathbf{w}}$ will be dominated by unbounded noises. When $\tau_\infty^{\mathbf{w}} = \infty$, the asymptotic behavior of $\hat{A}^{\mathbf{w}}$ will be exactly the same as $\bar{A}^{\mathbf{w}}$. When $0 < \tau_\infty^{\mathbf{w}} < \infty$, [23] provides us with the necessary tool to determine the spectrum distribution of $\hat{A}^{\mathbf{w}}$, which is the summation of a low-rank matrix $\bar{A}^{\mathbf{w}}$ and a noise matrix $\tilde{A}^{\mathbf{w}}$ whose eigenvalues are both constant scale.

For a symmetric matrix $X_n \in \mathbb{R}^{n\times n}$ with ordered eigenvalues $\lambda_1(X_n) \geq \cdots \geq \lambda_n(X_n)$. Let $\hat{f}_{X_n}$ be the empirical eigenvalue distribution, which is

$$f_{X_n} = \frac{1}{n}\sum_{j=1}^{n}\delta_{\lambda_j(X_n)}.$$

If f_{X_n} converges almost surely weakly, as $n \to \infty$, to a non-random compactly supported probability measure, we let f_{X_∞} denote the limit. Let a_{X_∞} and b_{X_∞} be the infimum and supremum of the support of f_{X_∞}, then the smallest and largest eigenvalue of X_n converge almost surely to a_{X_∞} and b_{X_∞}.

Since $\bar{A}^{\mathbf{w}} \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix with rank K, and $\lim_{n\to\infty}\lambda_i(\bar{A}^{\mathbf{w}}) = 1$ for $i = 2, \cdots, K$. Applying Theorem 2.1 and Remark 2.15 in [23], we have the following theorem.

Theorem 2.14. *The extreme eigenvalues of $\hat{A}^{\mathbf{w}}$ exhibit the following behavior as $n \to \infty$. We have* *that for each* $1 \leq i \leq K$,

$$\lambda_i(\hat{A}^{\mathbf{w}}) \xrightarrow{a.s.} \begin{cases} G_{f_X}^{-1}(1) & \text{if } G_{f_{X_\infty}}(b^+) < 1, \\ b & \text{otherwise}, \end{cases} \tag{2.2}$$

while for each fixed $i > K$, $\lambda_i(\hat{A}^{\mathbf{w}}) \xrightarrow{a.s.} b$. *Here,*

$$G_{f_X}(z) = \int \frac{1}{z-t} df_X(t) \text{ for } z \notin \text{supp } f_X,$$

is the Cauchy transform of f_X, $G_{f_X}^{-1}(\cdot)$ *is its functional inverse so that* $1/\pm\infty$ *stands for* 0.

By semicircle law [15], $f_{\tilde{A}^{\mathbf{w}}_\infty}(t) = 2\pi^{-1}\tau_\infty^{\mathbf{w}2}\{(1/\tau_\infty^{\mathbf{w}2} - t^2)_+\}^{1/2}$, then it's easy to verify that

$$G_{f_X}(z) = \begin{cases} \tau_\infty^{\mathbf{w}2}\{2z - 2(z^2 - 1/\tau_\infty^{\mathbf{w}2})^{1/2}\} & \text{if } z > 1/\tau_\infty^{\mathbf{w}}, \\ \tau_\infty^{\mathbf{w}2}\{2z + 2(z^2 - 1/\tau_\infty^{\mathbf{w}2})^{1/2}\} & \text{if } z < -1/\tau_\infty^{\mathbf{w}}, \end{cases}$$

$G_{f_X}(\pm 1/\tau_\infty^{\mathbf{w}}) = \pm 1/2\tau_\infty^{\mathbf{w}}$ and $G_{f_X}^{-1}(1/x) = (4x\tau_\infty^{\mathbf{w}2})^{-1} + x$ for $x \neq 0$. Thus, for $2 \leq i \leq K$,

$$\lambda_i(\hat{A}^{\mathbf{w}}) \xrightarrow{a.s.} \begin{cases} 1 + (4\tau_\infty^{\mathbf{w}2})^{-1}, & \text{if } \tau_\infty^{\mathbf{w}} > 1/2, \\ 1/\tau_\infty^{\mathbf{w}}, & \text{otherwise.} \end{cases}$$

□

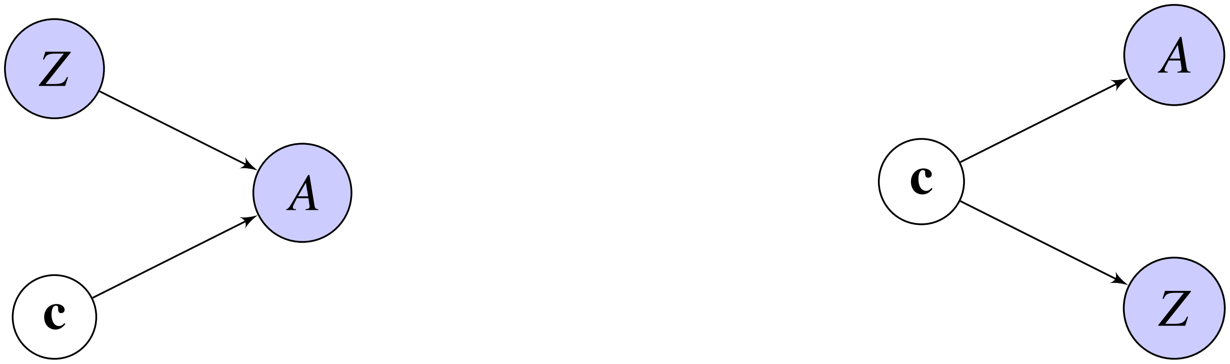
Chapter 3: Pairwise Covariate-Adjusted Block Model

3.1 Introduction

In the real world, the connection of nodes may depend on not only community structure but also on some nodal information. For example, in an ecological network, the predator-prey link between species may depend on their prey types as well as their habits, body sizes and living environment. Incorporating nodal information into the network model should help us recover a more accurate community structure.

Depending on the relationship between communities and covariates, there are in general two classes of models as shown in Figure 3.1: covariates-adjusted and covariates-confounding. $\mathbf{c}$, X and A respectively stands for latent community label, nodal information and adjacency matrix. In Figure 3.1a, the latent community and the covariates jointly determine the network structure. One typical example of this model is the friendship network between students. Students become friends for various reasons: they are in the same class; they have the same hobbies; they are of the same ethnic group. Without adjusting those covariates, it is hard to believe A represents any single community membership. We will analyze one such example in detail in Section 3.5. On the other hand, covariates sometimes carry the same community information as the adjacency matrix, which is shown in Figure 3.1b. The name 'confounding' comes from graph model [68]. Citation network is a perfect example of this model [127]. When the research topic is treated as the community label for each article, the citation links would largely depend on the research topics of the article pair. Meanwhile, the distribution of the keywords is also likely to be driven by the specific topic the article is about.

Most researchers modify SBM in the above two ways to incorporate covariates' information. For the covariates-adjusted model, [111] uses covariates to construct the prior for community label



(a) Covariates-adjusted

(b) Covariates-confounding

Figure 3.1: Different network models including covariates

and then generates edges by degree-corrected model; [144] proposes a directed network model with logistic function, but it does not consider possible community structure. For the covariates-confounding model, [139] uses a logistic model as the prior for community labels. [145] proposes a joint community detection criterion, which is an analog of modularity, to incorporate node features. [98] presents algorithms for two special classes of the latent space model that incorporate edge covariates. [142] proposes a generalized linear model with low-rank effects to model network edges, which implies the community structure though not mentioned explicitly.

In this chapter, we propose a simple yet effective model called PCABM, which extends the SBM by adjusting the probability of connections with the pairwise covariates. Through this model, we can learn how each covariate affects the connections by looking at its corresponding regression coefficient. Also, we show the consistency and asymptotic normality for MLE. Besides likelihood methods, we also propose a novel spectral clustering method called SCWA. Note that [27] also uses a modified version of spectral clustering to incorporate nodal covariates, but it is not based on a specific model. We prove desirable theoretical properties for SCWA applied to PCABM, and show that as a fast algorithm, using it as an initial estimator for the likelihood method usually leads to more accurate community detection than random initialization.

The rest of this chapter is organized as follows. In Section 3.2, we introduce the PCABM. We then show the asymptotic properties of the coefficient estimates as well as the community detection consistency in Section 3.3.1. Section 3.3.2 introduces SCWA and its asymptotic properties. Simulations and applications on real networks will be discussed in Section 3.4 and 3.5. We conclude this chapter with a short discussion in Section 3.6. All proofs are relegated to Section 3.7.

3.2 PCABM setup

Upon classical SBM, we assume in addition to A, we have additionally observed a pairwise p-dimensional vector $\mathbf{z}_{ij}$ between node i and j. Denote the collection of the pairwise covariate among nodes as $Z = [\mathbf{z}_{ij}^T] \in \mathbb{R}^{n^2 \times p}$. Define $\boldsymbol{\gamma}$ as a fixed common coefficient vector for all node pairs (i, j) with the true value denoted by $\boldsymbol{\gamma}^0$. For $i < j$, conditioned on $\mathbf{c}, Z, \boldsymbol{\gamma}^0, B$, A_{ij}'s are independent and

$$A_{ij} \sim \text{Poisson}(\lambda_{ij}),\ \lambda_{ij} = B_{c_i c_j} e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}.$$

In this thesis, we only consider sparse networks, which means λ_{ij} is far less than 1. For simplicity, we assume $\lambda_{ij} < 1/2$. The specific term $e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}$ is introduced here to adjust the SBM with the pairwise covariates. φ_n is no longer the expected degree as in [147], but we can still use it to measure the rate of ρ_n. For theoretical purposes, we need the following conditions on Z and $\boldsymbol{\gamma}^0$.

Condition 3.1. *$\{\mathbf{z}_{ij}, i < j\}$ are i.i.d. and $\mathbf{z}_{ij} = \mathbf{z}_{ji}$.*

Condition 3.2. *$\{\mathbf{z}_{ij}, i < j\}$ are uniformly bounded, i.e., for $\forall i < j$, $\|\mathbf{z}_{ij}\|_\infty \le \zeta$, where $\zeta > 0$ is some constant.*

Remark 3.1. *The bounded support condition for z_{ij} is introduced to simply our proofs. It could be relaxed to allow the upper bound grow slowly with network size n.*

By Condition 3.2, we know that $e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}$ is also uniformly bounded. We define the bound as $\beta_l \le e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0} \le \beta_u$. Thus, the following expectations exist: $\theta(\boldsymbol{\gamma}^0) \equiv \mathbb{E} e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0} \in \mathbb{R}^+$, $\boldsymbol{\mu}(\boldsymbol{\gamma}^0) \equiv \mathbb{E} \mathbf{z}_{ij} e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0} \in \mathbb{R}^p$, and $\Sigma(\boldsymbol{\gamma}^0) \equiv \mathbb{E} \mathbf{z}_{ij} \mathbf{z}_{ij}^T e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0} \in \mathbb{R}^{p \times p}$.

Condition 3.3. *$\Sigma(\boldsymbol{\gamma}^0) - \theta(\boldsymbol{\gamma}^0)^{-1} \boldsymbol{\mu}(\boldsymbol{\gamma}^0)^{\otimes 2}$ is positive definite.*

Remark 3.2. *Condition 3.3 is imposed to ensure that $\boldsymbol{\gamma}^0$ is the unique solution to maximize the likelihood in the population version. Consider the function $g(\boldsymbol{\gamma}) = \theta(\boldsymbol{\gamma}) \Sigma(\boldsymbol{\gamma}) - \boldsymbol{\mu}(\boldsymbol{\gamma})^{\otimes 2}$. If $\boldsymbol{\gamma}^0 = \mathbf{0}$,* *we have* $g(\mathbf{0}) = \mathbb{E}[\mathbf{z}^{\otimes 2}] - \mathbb{E}[\mathbf{z}]^{\otimes 2} = \mathrm{cov}(\mathbf{z})$. *To avoid multicollinearity, it's natural for us to require* $\mathrm{cov}(\mathbf{z})$ *to be positive definite. Similarly, we require* $g(\boldsymbol{\gamma}^{\mathbf{0}})$ *to be positive definite at the true value* $\boldsymbol{\gamma}^{\mathbf{0}}$.

Define

$$n_k(\mathbf{e}) \equiv \sum_{i=1}^{n} \mathbb{1}_{e_i=k},\ O_{kl}(\mathbf{e}) \equiv \sum_{ij} A_{ij}\mathbb{1}_{e_i=k,e_j=l},\ E_{kl}(\mathbf{e},\boldsymbol{\gamma}) \equiv \sum_{i\neq j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}}\mathbb{1}_{e_i=k,e_j=l} = \sum_{(i,j)\in s_{\mathbf{e}}(k,l)} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}},$$

where $s_{\mathbf{e}}(k,l) = \{(i,j)|e_i = k, e_j = l, i \neq j\}$. Under assignment $\mathbf{e}$, $n_k(\mathbf{e})$ represents the number of nodes in community k. For $k \neq l$, O_{kl} is the total number of edges between communities k and l; for $k = l$, O_{kk} is twice of the number of edges within community k. E_{kl} is the summation of all pair-level factors. We can write the log-likelihood function as

$$\log \mathcal{L}(\mathbf{e},\boldsymbol{\gamma},B,\boldsymbol{\pi}|A,Z) \propto \sum_k n_k(\mathbf{e})\log\pi_k + \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e})\log B_{kl} - \frac{1}{2}\sum_{kl} B_{kl}E_{kl}(\mathbf{e},\boldsymbol{\gamma}) + \sum_{i<j} A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma}.$$

For fixed $\mathbf{e}$ and $\boldsymbol{\gamma}$, the MLE of $\boldsymbol{\pi}$ and B is $\hat{\pi}_k(\mathbf{e}) = n_k(\mathbf{e})/n$ and $\hat{B}_{kl}(\mathbf{e}) = O_{kl}(\mathbf{e})/E_{kl}(\mathbf{e},\boldsymbol{\gamma})$. Plugging $\hat{B}(\mathbf{e}), \hat{\boldsymbol{\pi}}(\mathbf{e})$ into the original log-likelihood and discarding the constant terms, we have

$$\log \mathcal{L}(\mathbf{e},\boldsymbol{\gamma},\hat{B},\hat{\boldsymbol{\pi}}|A,Z) \propto \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e})\log\frac{O_{kl}(\mathbf{e})}{E_{kl}(\mathbf{e},\boldsymbol{\gamma})} + \sum_{i<j} A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma} + \sum_k n_k(\mathbf{e})\log\frac{n_k(\mathbf{e})}{n}. \tag{3.1}$$

Out target is to maximize (3.1) w.r.t. $\mathbf{e}$ and $\boldsymbol{\gamma}$. To achieve that, we apply a two-step procedure here, which means we estimate $\boldsymbol{\gamma}^0$ and $\mathbf{c}$ sequentially. The theoretical properties for the MLE $\hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{c}}$ will be demonstrated in the following section.

3.3 Theory and algorithms for PCABM

In this section, we will introduce two approaches for PCABM estimation. One is likelihood method in Section 3.3.1, and the other is a modified spectral clustering method in Section 3.3.2. While the former is the canonical analysis for a new model, the latter is a more efficient approach.

3.3.1 Likelihood method

The likelihood theory will be divided into three parts. First, we investigate the property of MLE $\hat{\boldsymbol{\gamma}}(\mathbf{e})$ under some random community assignment $\mathbf{e}$. We derive the consistency and asymptotic normality of $\hat{\boldsymbol{\gamma}}(\mathbf{e})$ under some mild conditions of $\mathbf{e}$. Second, we demonstrate that under a large class of criteria Q, given $\hat{\boldsymbol{\gamma}}(\mathbf{e})$ is consistent, the consistency of $\hat{\mathbf{c}}(\mathbf{e})$ is guaranteed under PCABM. At last, we show that under our likelihood criterion, $\hat{\mathbf{c}}$ is consistent under any $\hat{\boldsymbol{\gamma}}$, even if it's not consistent, which could potentially save time by skipping the $\hat{\boldsymbol{\gamma}}$ estimation step.

A. Consistency of $\hat{\boldsymbol{\gamma}}$

Consider the terms in (3.1) only containing $\boldsymbol{\gamma}$, which is

$$\ell_{\mathbf{e}}(\boldsymbol{\gamma}) \equiv \sum_{i<j} A_{ij}\mathbf{z}_{ij}^T\boldsymbol{\gamma} - \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e}) \log E_{kl}(\mathbf{e},\boldsymbol{\gamma}).$$

Define $\hat{\boldsymbol{\gamma}}(\mathbf{e}) \equiv \arg\max_{\gamma} \ell_{\mathbf{e}}(\boldsymbol{\gamma})$ to be the MLE. When there is no need to emphasize, we will simply omit $\mathbf{e}$. To derive the asymptotic property for $\hat{\boldsymbol{\gamma}}$, the following regularity condition regarding the community assignment $\mathbf{e}$ is assumed.

Condition 3.4 (non-degeneracy)**.** *For any $k \in [K]$, $\kappa_1 \le |n_k(\mathbf{e})/n| \le \kappa_2$, where $0 < \kappa_1 \le \kappa_2 < 1$ are some positive constants.*

Condition 3.4 is very mild and natural since we know the size of each community is approximately a constant proportion of the total number of nodes under the true model. The consistency of $\hat{\boldsymbol{\gamma}}$ is guaranteed by the following theorem.

Theorem 3.1. *Under PCABM, assume Condition 3.1, 3.3 and 3.4 hold. As $n \to \infty$, if $N_n\rho_n \to \infty$, $\hat{\boldsymbol{\gamma}}(\mathbf{e}) \xrightarrow{p} \boldsymbol{\gamma}^0$.*

This theorem saves us from tedious estimations of $\boldsymbol{\gamma}^0$ and $\mathbf{c}$ iteratively. It shows that we can get a consistent estimate of $\boldsymbol{\gamma}^0$ under any community assignment $\mathbf{e}$ with mild conditions. With Theorem 3.1, we no longer need a good initial estimate of $\mathbf{c}$ to guarantee the consistency of $\hat{\boldsymbol{\gamma}}$. We can simply take any random non-degenerate assignment $\mathbf{e}$ as the initial class label and optimize the likelihood function to get $\hat{\boldsymbol{\gamma}}(\mathbf{e})$. Next, we present the asymptotic normality property for $\hat{\boldsymbol{\gamma}}(\mathbf{e})$.

Theorem 3.2. *Under PCABM, assume Condition 3.1, 3.3 and 3.4 hold. As $n \to \infty$, if $N_n\rho_n \to \infty$, the asymptotic distribution of $\sqrt{N_n\rho_n}(\hat{\boldsymbol{\gamma}}(\mathbf{e})-\boldsymbol{\gamma}^0)$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\Sigma_\infty^{-1}(\boldsymbol{\gamma}^0)$, where $\Sigma_\infty(\boldsymbol{\gamma}^0) = \sum_{ab} \Omega_{ab}\pi_a\pi_b(\Sigma(\boldsymbol{\gamma}^0) - \theta(\boldsymbol{\gamma}^0)^{-1}\boldsymbol{\mu}(\boldsymbol{\gamma}^0)^{\otimes 2})$.*

Different from [144], in which the network is dense, the convergence rate in our theorem is $\sqrt{N_n\rho_n}$ rather than $\sqrt{N_n}$ since the effective number of edges is reduced from N_n to $N_n\rho_n$. The asymptotic covariance matrix $\Sigma_\infty^{-1}(\boldsymbol{\gamma}^0)$ depends on $\theta(\boldsymbol{\gamma}^0), \boldsymbol{\mu}(\boldsymbol{\gamma}^0), \Sigma(\boldsymbol{\gamma}^0)$, which can be estimated empirically.

B. Consistency of $\hat{\mathbf{c}}$

The second step is to optimize (3.1) w.r.t. $\mathbf{e}$ by plugging $\hat{\boldsymbol{\gamma}}$ from the first step. The log-likelihood function to maximize is defined as

$$\ell_{\hat{\boldsymbol{\gamma}}}(\mathbf{e}) = \frac{1}{2}\sum_{kl} O_{kl}(\mathbf{e}) \log \frac{O_{kl}(\mathbf{e})}{E_{kl}(\mathbf{e}, \hat{\boldsymbol{\gamma}})} + \sum_k n_k(\mathbf{e}) \log \frac{n_k(\mathbf{e})}{n}. \tag{3.2}$$

Instead of dealing with $\ell_{\hat{\boldsymbol{\gamma}}}(\mathbf{e})$ directly, as in [147], we first show the consistency of $\hat{\mathbf{c}}$ under a large class of criteria defined as

$$Q(\mathbf{e}, \hat{\boldsymbol{\gamma}}) = F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e}, \hat{\boldsymbol{\gamma}})}{2N_n}\right). \tag{3.3}$$

Next, we show our likelihood method falls within this class of criteria, so the consistency of label estimation naturally follows. We say the criterion Q is consistent if the labels obtained by maximizing the criterion $\hat{\mathbf{c}} \equiv \arg\max_{\mathbf{e}} Q(\mathbf{e}, \hat{\boldsymbol{\gamma}})$ is consistent.

Here, we consider two versions of consistency. While weak consistency means convergence in probability: $\Pr(\sum_{i=1}^n \mathbb{1}_{\hat{c}_i \neq c_i}/n < \varepsilon) \to 1$ for any $\varepsilon > 0$ as $n \to \infty$, strong consistency means convergence almost surely: $\Pr(\hat{\mathbf{c}} \neq \mathbf{c}) \to 1$, as $n \to \infty$. Here, the equality is interpreted to mean membership in the same equivalence class with respect to label permutations. We prove both versions of consistency under the same sparsity conditions as those in [147], respectively.

Naturally, one key condition of Q is the maximum value under the 'population version' should be obtained at the true community assignment. Given a community assignment $\mathbf{e} \in [K]^n$, we define $R(\mathbf{e}) \in \mathbb{R}^{K\times K}$ with its elements being $R_{ka}(\mathbf{e}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{e_i=k,c_i=a}$. One can view R as the empirical joint distribution of $\mathbf{e}$ and $\mathbf{c}$. The condition for maximum value and some other regularity conditions satisfied by F are as follows, which are analog to those in [147].

Condition 3.5. *$F(R\Omega R^T, RJR^T)$ is uniquely maximized over $\mathcal{R} = \{R : R \geq 0, R^T\mathbf{1} = \boldsymbol{\pi}\}$ by $R = D(\boldsymbol{\pi})$.*

Condition 3.6. *Some regularity conditions hold for F.*

1. *F is Lipschitz in its arguments and $F(cX_0, cY_0) = cF(X_0, Y_0)$ for constant $c \neq 0$.*
2. *The directional derivatives $\frac{\partial^2 F}{\partial \epsilon^2}(X_0+\epsilon(X_1-X_0), Y_0+\epsilon(Y_1-Y_0))|_{\epsilon=0+}$ are continuous in (X_1, Y_1) for all (X_0, Y_0) in a neighborhood of $(D(\boldsymbol{\pi})\Omega D(\boldsymbol{\pi})^T, \boldsymbol{\pi}\boldsymbol{\pi}^T)$.*
3. *Let $G(R,\Omega) = F(R\Omega R^T, RJR^T)$. On $\mathcal{R}$, the gradient $\frac{\partial G((1-\epsilon)D(\boldsymbol{\pi})+\epsilon R,\Omega)}{\partial \epsilon}|_{\epsilon=0+} < -C < 0$ for all $\boldsymbol{\pi}$ and Ω.*

Now the main theorem can be stated as follows.

Theorem 3.3. *For all criteria Q of the form (3.3), if $\boldsymbol{\pi}, \Omega, F$ satisfy Condition 3.5 and 3.6, and $\hat{\boldsymbol{\gamma}}$ is consistent, then under PCABM, Q is weakly consistent if $\varphi_n \longrightarrow \infty$ and strongly consistent if $\varphi_n/\log n \longrightarrow \infty$.*

Remark 3.3. *Consider a network generated by PCABM. If we ignore the covariates and use SBM instead, it is equivalent to estimate PCABM with $\hat{\boldsymbol{\gamma}} = 0$. Suppose nodes i, j and k belong to the first community. We have $\lambda_{ik} = B_{11}e^{\mathbf{z}_{ik}^T\boldsymbol{\gamma}^0} \neq B_{11}e^{\mathbf{z}_{jk}^T\boldsymbol{\gamma}^0} = \lambda_{jk}$ unless $\mathbf{z}_{ik}^T\boldsymbol{\gamma}^0 = \mathbf{z}_{jk}^T\boldsymbol{\gamma}^0$. Therefore, the nodes in the same community could behave very differently depending on the value of the pairwise covariates. As a result, we would expect the regular SBM to have difficulty in detecting* *the communities when the magnitude of $\boldsymbol{\gamma}^0$ is large. We will demonstrate this point through a simulation example in Section 3.4 where we vary the magnitude of $\boldsymbol{\gamma}^0$.*

We will show in Section 3.7.4 that the log likelihood satisfy above conditions, thus is consistent. Surprisingly, we further show that the consistency of $\hat{\boldsymbol{\gamma}}$ is not necessary to guarantee the consistency of $\hat{\mathbf{c}}$, which will save us from parameter estimation. Under the PCABM, we could simply plug $\hat{\boldsymbol{\gamma}} = \mathbf{0}$ into the likelihood and obtain $\hat{\mathbf{c}}$. The theorem is formally stated as follows.

Corollary 3.4. *Under PCABM, the likelihood criterion (3.2) is weakly consistent if $\varphi_n \to \infty$ and strongly consistent if $\varphi_n / \log n \to \infty$ for any $\hat{\boldsymbol{\gamma}}$.*

At the first glance, Theorem 3.4 may be kind of astonishing. It shows the community block probability and the covariates probability are decoupled and we could estimate them separately. One way to understand this is to see the covariates part from an expectation view, then you may still get a consistent estimate $\hat{\mathbf{c}}$ asymptotically. However, in practice, we always deal with finite samples, which might cause large bias by simply ignoring the covariates.

Since finding $\hat{\mathbf{c}}$ is a non-convex problem, we use tabu search [20, 66] to find a solution. The idea of is to randomly switch the class label for a pair of nodes. If the value of log-likelihood function increases after switching, we proceed with the switch. Otherwise, we ignore the switch by sticking with old labels. Because this algorithm is greedy, to avoid being trapped in local maximum, we 'tabu' those nodes whose labels have been switched in a preceding period time, i.e. we don't consider switching the label of a node if it is in the tabu set. Though global maximum is not guaranteed, tabu search usually gives satisfactory results in our numerical experience. The detailed algorithm is described as follows. Although theoretically we don't need a consistent $\hat{\boldsymbol{\gamma}}$ to guarantee the consistency of $\hat{\mathbf{c}}$, for the reasons we have disscused, we still use a two step procedure to get more stable results.

Algorithm 3.1 PCABM.MLE0

```
Input: adjacency matrix $A$, pairwise covariates $Z$, initial assignment $\mathbf{e}$, likelihood function $\mathcal{L}(\mathbf{e}, \boldsymbol{\gamma}, \hat{B}, \hat{\boldsymbol{\pi}})$ and number of communities $K$
Output: coefficient estimate $\hat{\boldsymbol{\gamma}}$ and community estimate $\hat{\mathbf{c}}$
1: Optimize $\ell_{\mathbf{e}}(\boldsymbol{\gamma})$ by some optimization algorithm (e.g., BFGS) to derive $\hat{\boldsymbol{\gamma}}$.
2: Use tabu search to optimize $\ell_{\hat{\boldsymbol{\gamma}}}(\mathbf{e})$ to get $\hat{\mathbf{c}}$.
```

3.3.2 Spectral clustering with adjustment

Though the likelihood method has appealing theoretical properties, tabu search can sometimes be slow when the network size is large. Also, the community detection results can be sensitive to the initial label assignments **e**. As a result, we aim to propose a computationally efficient algorithm in the flavor of spectral clustering [119] to be used as the initial solution for PCABM.

A. $(1+\epsilon)$ **spectral clustering algorithm**

From now on, we use PCABM$(M, B, Z, \boldsymbol{\gamma}^0)$ to represent PCABM generated with parameters in the parentheses. Let $G_k = G_k(M) = \{1 \le i \le n : c_i = k\}$, then $n_k = |G_k|$ for $k = 1, \cdots, K$. Let $n_{\min} = \min_{1\le k\le K} n_k$, $n_{\max} = \max_{1\le k\le K} n_k$ and $n'_{\max}$ is the second largest community size. We define k-means clustering in another form

$$(\hat{M}, \hat{X}) = \arg\min_{M\in\mathcal{M}_{n,K}, X\in\mathbb{R}^{K\times K}} \|MX - U_A\|_F^2. \tag{3.4}$$

Though finding a global minimizer for the k-means problem (3.4) is NP-hard [11], for any positive constant ϵ, we have efficient algorithms to find an $(1+\epsilon)$-approximate solution [86, 96]

$$\begin{aligned} &(\hat{M}, \hat{X}) \in \mathcal{M}_{n,K} \times \mathbb{R}^{K\times K} \\ s.t. \qquad &\|\hat{M}\hat{X} - U_A\|_F^2 \le (1+\epsilon) \min_{M\in\mathcal{M}_{n,K}, X\in\mathbb{R}^{K\times K}} \|MX - U_A\|_F^2. \end{aligned}$$

The goal of community detection is to find $\hat{M}$ that is close to M. To define a loss function, we need to take permutation into account. Let $\mathcal{S}_K$ be the space of all $K \times K$ permutation matrices. Following [91], we define two measures of estimation error: the overall relative error and the worst case relative error:

$$L_1(\hat{M}, M) = n^{-1} \min_{S \in \mathcal{S}_K} \|\hat{M}S - M\|_0,$$

$$L_2(\hat{M}, M) = \min_{S \in \mathcal{S}_K} \max_{1 \le k \le K} n_k^{-1} \|(\hat{M}S)_{G_k \cdot} - M_{G_k \cdot}\|_0.$$

It can be seen that $0 \le L_1(\hat{M}, M) \le L_2(\hat{M}, M) \le 2$. While L_1 measures the overall proportion of mis-clustered nodes, L_2 measures the worst performance across all communities.

B. Spectral clustering with adjustment

The existence of covariates in PCABM prevents us from applying spectral clustering directly on A. Unlike SBM where A is generated from a low rank matrix P, A consists of both community and covariate information in PCABM. Since $P_{ij} = \mathbb{E}[A_{ij}/e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}]$, one natural idea to take advantage of the low rank structure is to remove the covariate effects, i.e. using the adjusted adjacency matrix $[A_{ij}/e^{\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}]$ for spectral clustering. However, in practice, we don't know what the the true parameter $\boldsymbol{\gamma}^0$ is, so we need to replace $\boldsymbol{\gamma}^0$ with empirical estimate $\hat{\boldsymbol{\gamma}}$. Define the adjusted adjacency matrix as $A' = [A'_{ij}]$ where $A'_{ij} = A_{ij} \exp(-\mathbf{z}_{ij}^T \hat{\boldsymbol{\gamma}})$. By the consistency of $\hat{\boldsymbol{\gamma}}$ proved in Theorem 3.1, we hope $\|A' - P\|$ can be bounded in probability as in [91]. Based on this bound, we could then apply the normal spectral clustering algorithm on matrix A' to detect the community. This SCWA algorithm is elaborated in Algorithm 3.2. Also, the following regularity conditions are assumed throughout this section, which is similar to the degree sparsity condition in [91].

Condition 3.7. $\|\Omega\|_{\max} \le 1$ *and* $\varphi_n \ge C \log n$ *for some constant* $C > 0$.

Theorem 3.5 (Spectral bound of Poisson random matrices)**.** *Let A be the adjacency matrix generated by PCABM* $(M, B, Z, \boldsymbol{\gamma}^0)$*, and the adjusted adjacency matrix* A' *is derived from the MLE* $\hat{\boldsymbol{\gamma}}$*. Assume Condition 3.1 3.2 3.4 3.7 hold. For any constant* $r > 0$*, there exists a constant C such that* $\|A' - P\| \le C\sqrt{\varphi_n}$ *with probability at least* $1 - n^{-r}$*.*

Algorithm 3.2 PCABM.SCWA

```
Input: adjacency matrix $A$, pairwise covariates $Z$, initial assignment $\mathbf{e}$, likelihood function
    $\mathcal{L}(\mathbf{e}, \boldsymbol{\gamma}, \hat{B}, \hat{\boldsymbol{\pi}})$, number of communities $K$ and approximation parameter $\epsilon$
Output: coefficient estimate $\hat{\boldsymbol{\gamma}}$ and community estimate $\hat{\mathbf{c}}$
1: Optimize $\ell_{\mathbf{e}}(\boldsymbol{\gamma})$ by some optimization algorithm (e.g., BFGS) to derive $\hat{\boldsymbol{\gamma}}$.
2: Divide $A_{ij}$ by $\exp(\mathbf{z}_{ij}^T \hat{\boldsymbol{\gamma}})$ to get $A'_{ij}$.
3: Calculate $U_{A'} \in \mathbb{R}^{n \times K}$ consisting of the leading $K$ eigenvectors (ordered in absolute eigen-
    value) of $A'$.
4: Calculate the $(1+\epsilon)$-approximate solution $(\hat{M}, \hat{X})$ to the $k$-means problem (3.4) with $K$ clusters
    and input matrix $U_{A'}$.
5: Output $\hat{\mathbf{c}}$ according to $\hat{M}$.
```

With similar proof of Theorem 3.1 in [91], we can prove the following Theorem 3.6 by combining Lemmas 5.1 and 5.3 in [91], and Theorem 3.5.

Theorem 3.6. *Based on the conditions of Theorem 3.5, assume that $P = MBM^T$ is of rank K with the smallest absolute non-zero eigenvalue at least ξ_n. Let $\hat{M}$ be the output of spectral clustering using $(1+\epsilon)$ approximate k-means on A'. For any constant $r > 0$, there exists an absolute constant $C > 0$, such that, if*

$$(2+\epsilon)\frac{Kn\rho_n}{\xi_n^2} < C,$$

then, with probability at least $1 - n^{-r}$, there exist subsets $H_k \subset G_k$ for $k = 1, \cdots, K$, and a $K \times K$ permutation matrix J such that $\hat{M}_{G\cdot} J = M_{G\cdot}$, where $G = \cup_{k=1}^{K}(G_k \setminus H_k)$, and

$$\sum_{k=1}^{K} \frac{|H_k|}{n_k} \le C^{-1}(2+\epsilon)\frac{Kn\rho_n}{\xi_n^2}. \tag{3.5}$$

Inequality (3.5) provides an error bound for $L_2(\hat{M}, M)$. In set H_k, the clustering accuracy of nodes can not be guaranteed. Theorem 3.6 doesn't provide us with an error bound in a straightforward form since ξ_n contains ρ_n. The following corollary gives us a clearer view of the error bound in terms of model parameters.

Corollary 3.7. *Based on the conditions of Theorem 3.5, assume Ω's minimum absolute eigenvalue bounded below by $\tau > 0$ and $\|\Omega\|_{\max} = 1$. Let $\hat{M}$ be the output of spectral clustering using $(1+\epsilon)$*

approximate k-means. For any constant $r > 0$, there exists an absolute constant C such that if

$$(2+\epsilon)\frac{Kn}{n_{\min}^2\tau^2\rho_n} < C,$$

then with probability at least $1 - n^{-r}$,

$$L_2(\hat{M}, M) \le C^{-1}(2+\epsilon)\frac{Kn}{n_{\min}^2\tau^2\rho_n} \text{ and } L_1(\hat{M}, M) \le C^{-1}(2+\epsilon)\frac{Kn'_{\max}}{n_{\min}^2\tau^2\rho_n}.$$

The result provides us with the same convergence rate as in [91]. The intuition is that if we see the covariate from an expectation view, then it's just a constant multiplied to the original probability, the rate of convergence shouldn't be affected.

Algorithm 3.3 PCABM.MLE

```
Input: adjacency matrix $A$, pairwise covariates $Z$, initial classes $\mathbf{e}$, likelihood function
    $\mathcal{L}(\mathbf{e}, \boldsymbol{\gamma}, \hat{B}, \hat{\boldsymbol{\pi}})$, number of communities $K$ and approximation parameter $\epsilon$
Output: coefficient estimate $\hat{\boldsymbol{\gamma}}$ and community estimate $\hat{\mathbf{c}}$
  Use Algorithm 3.2 to get an initial community estimate $\tilde{\mathbf{c}}$.
  Use $\tilde{\mathbf{c}}$ as the initial value in $\ell_{\hat{\boldsymbol{\gamma}}}(\tilde{\mathbf{c}})$ for tabu search to derive $\hat{\mathbf{c}}$.
```

Compared with SCWA, the likelihood tabu search usually leads to more precise results but takes longer time. Also, the results of the likelihood tabu search are sensitive to the initial labels **e** in some settings. To combine the advantages of those two methods, we propose to use the results of SCWA as the initial solution for tabu search (PCABM.MLE as described in Algorithm 3.3), which shows better empirical performance than each method alone. We will conduct extensive simulation studies in Section 3.4.

3.4 Simulations

For all simulations, we consider $K = 2$ communities with prior probability $\pi_1 = \pi_2 = 0.5$. In addition, we fix $\Omega = \left(\begin{smallmatrix} 2 & 1 \\ 1 & 2 \end{smallmatrix}\right)$ and generate data by applying the following procedure:

1. Determine parameters ρ_n and $\boldsymbol{\gamma}^0$. Generate Z from certain distributions.

2. Generate adjacency matrix $A = [A_{ij}]$ by PCABM with parameters in step 1.

We use the same pairwise covariate Z in all simulations, whose 5 entries are generated independently from Bernoulli(0.1), Poisson(0.1), Uniform$[0,1]$, Exponential(0.3), $N(0,0.3)$, respectively. We use a uniform bound 2 for Z. The parameters for each distribution are chosen to make the variances of covariates similar. In the following experiments, we will omit the above procedures and only state the other parameters.

3.4.1 Coefficient estimation

For PCABM, although we could skip the step of estimating $\boldsymbol{\gamma}$, but empirically a good estimation would provide more stable results. Also, we want to check the consistency of $\hat{\boldsymbol{\gamma}}$ in Section 3.3.1. We ran 100 simulations respectively for $n = 100, 200, 300, 400, 500$. The parameters are set as $\rho_n = 2(\log n)^{1.5}/n$, $\boldsymbol{\gamma}^0 = (0.4, 0.8, 1.2, 1.6, 2)^T$. The parameter ρ_n is chosen to satisfy the strong consistency condition.

We obtain $\hat{\boldsymbol{\gamma}}$ by using BFGS to optimize $\ell_{\mathbf{e}}(\boldsymbol{\gamma})$ assuming no community structure exists. We also tried to optimize under random community assignments, which yields similar results. This validates Theorem 3.1, showing that estimating $\boldsymbol{\gamma}$ and $\mathbf{c}$ are decoupled. We list the mean and standard deviation of $\hat{\boldsymbol{\gamma}}$ in Table 3.1. It is clear that $\hat{\boldsymbol{\gamma}}$ is very close to $\boldsymbol{\gamma}^0$ even for a small network. The shrinkage of standard deviation shows consistency of MLE.

By taking a closer look at the network of size $n = 500$, we compare the distribution of $\hat{\boldsymbol{\gamma}}$ with the theoretical asymptotic normal distribution derived in Theorem 3.2. We show the histogram for the first three coefficients in Figure 3.2. We can see that the empirical distribution matches well with the theoretical counterpart.

3.4.2 Community detection

After obtaining $\hat{\boldsymbol{\gamma}}$, we focus on the estimation of community labels. Under PCABM, there are three parameters that we could tune to change the property of the network: $\boldsymbol{\gamma}^0$, ρ_n and n. To

Table 3.1: Simulated results for distribution of $\hat{\boldsymbol{\gamma}}$, displayed as mean (standard deviation)

n	$\gamma_1^0 = 0.4$	$\gamma_2^0 = 0.8$	$\gamma_3^0 = 1.2$	$\gamma_4^0 = 1.6$	$\gamma_5^0 = 2$
100	0.4018 (0.0269)	0.8003 (0.0184)	1.1944 (0.0369)	1.6153 (0.0309)	2.0214 (0.0302)
200	0.3929 (0.0139)	0.7956 (0.0117)	1.1648 (0.0228)	1.5901 (0.0151)	2.0057 (0.0205)
300	0.4162 (0.0114)	0.7953 (0.0073)	1.1607 (0.0186)	1.5962 (0.0086)	1.9912 (0.0120)
400	0.3931 (0.0091)	0.8049 (0.0057)	1.1960 (0.0144)	1.6075 (0.0081)	2.0122 (0.0097)
500	0.3944 (0.0081)	0.7998 (0.0051)	1.2017 (0.0124)	1.5970 (0.0086)	2.0038 (0.0099)

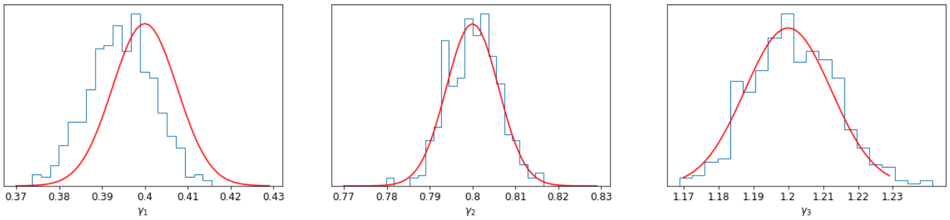


Figure 3.2: Simulation results for $\hat{\boldsymbol{\gamma}}$ compared with theoretical values

illustrate the impact of these parameters on the performance of community detection, we vary one parameter while fixing the remaining two in each experiment. More specifically, we consider the form $\rho_n = c_\rho(\log n)^{1.5}/n$ and $\boldsymbol{\gamma}^0 = c_\gamma(0.4, 0.8, 1.2, 1.6, 2)^T$ in which we will vary the multipliers c_ρ and c_γ. We also did one more experiment by plugging different $\hat{\boldsymbol{\gamma}}$ into $\ell_{\hat{\gamma}}(\mathbf{e})$ and estimate $\hat{\mathbf{c}}$. Here we use one dimensional covariate following $Poisson(0.1)$. The detailed parameter settings for the four experiments are as follows.

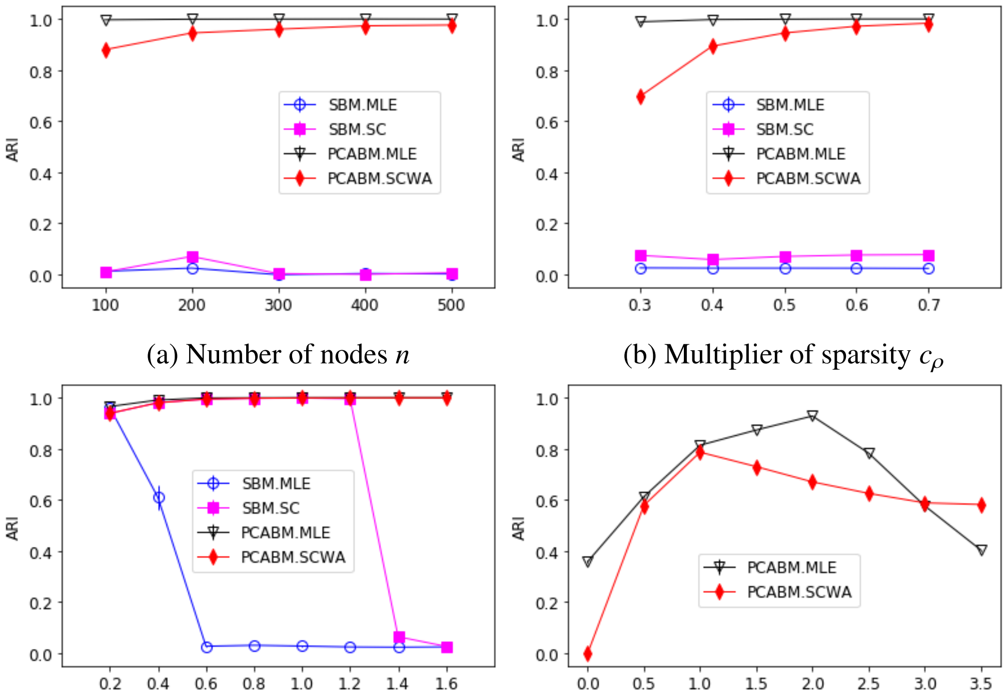
1. $n \in \{100, 200, 300, 400, 500\}$, with $c_\rho = 0.5$ and $c_\gamma = 1.4$.

2. $c_\rho \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, with $n = 200$ and $c_\gamma = 1.4$.

3. $c_\gamma \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6\}$, with $n = 200$ and $c_\rho = 1$.

4. $\hat{\gamma} \in \{0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$, with $\boldsymbol{\gamma}^0 = 2$, $n = 500$ and $c_\rho = 1$.

The results for the three experiments are presented in Figure 3.3. Each experiment is carried out for 100 times. The error rate is reported in terms of average ARI. SBM.MLE and SBM.SC refer to likelihood and spectral clustering methods under SBM, respectively. PCABM.MLE and PCABM.SCWA refer to Algorithms 3.3 and 3.2 respectively.



(a) Number of nodes n

(b) Multiplier of sparsity c_ρ

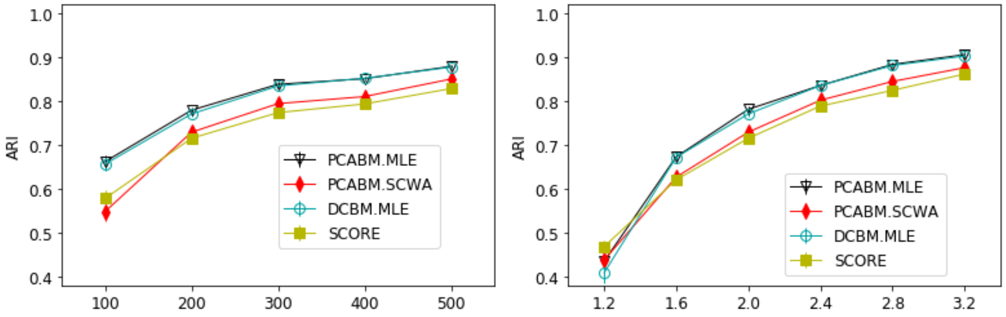
(c) Multiplier of coefficient c_γ

(d) Coefficient γ

Figure 3.3: Simulation results under PCABM for different parameter settings

When the number of nodes increases, it is clear from the first plot in Figure 3.3 that both algorithms under SBM provides us little more than random guess. On the other hand, both PCABM-based algorithms perform quite well with PCABM.MLE having nearly perfect community detection performance throughout all n. Note that SCWA gradually catches up with MLE as we have more nodes. We observe a similar phenomenon when the sparsity level is changed. When the scale of γ^0 is changed, both algorithms under PCABM still yield good results. As we know, when $\gamma^0 = \mathbf{0}$, our model reduces to SBM, so it is not surprising that SBM.MLE and SBM.SC both perform well when the magnitude of γ^0 is relatively small and fail when the magnitude increases. Also, it appears that compared with the likelihood method, spectral clustering is more robust to model misspecification. The last plot of Figure 3.3 shows what we claim earlier. Although Theorem 3.4 doesn't require a consistent $\hat{\gamma}$ to guarantee the consistency of $\hat{\mathbf{c}}$, using a more accurate $\hat{\gamma}$

would yield better result in finite sample experiment. We can also see the improvement of likelihood method upon SCWA when $\hat{\gamma}$ is accurate. When it's not, it's possible for likelihood method to introduce more bias.



(a) Number of nodes n with $c_\rho = 2$ (b) Multiplier of sparsity c_ρ with $n = 200$

Figure 3.4: Simulation results under DCBM for different parameter settings

To show the robustness of our algorithms, we also apply Algorithms 3.2 and 3.3 to networks generated by DCBM, which can be viewed as a misspecified case. The degree parameter for each node is chosen from $\{0.5, 1.5\}$ with equal probability, while other parameters follow the same scheme as PCABM. For covariates construction, we take $z_{ij} = \log d_i + \log d_j$, where d_i is the degree of node i. As a comparison, we also implemented the likelihood method in [147] and the SCORE method in [78]. In Figure 3.4, all algorithms perform better when the network is larger or denser. Surprisingly, SCWA is slightly better than SCORE in most cases. The two likelihood-based methods perform similarly and better than their spectral counterparts. The flexibility of PCABM allows us to model any factors that may contribute to the structure of network in addition to the underlying communities.

3.5 Real data examples

3.5.1 Example 1: political blogs

The first real world dataset we used is the network of political blogs created by [6]. The nodes are blogs about US politics and edges represent hyperlinks between them. We treated the network as undirected and only focus on the largest connected component of the network, resulting in a network with 1,222 nodes and 16,714 edges.

Because there are no other covariates available in this dataset, we created the pairwise covariates by aggregating the degree information. We let $z_{ij} = \log(d_i \times d_j)$, where d_i is the degree for the i-th node. The corresponding $\hat{\gamma}$ is 1.001, which essentially reduces PCABM to DCBM. Table 3.2 summarizes the performance comparison of PCABM with some existing results on this dataset. Besides ARI, we also evaluated normalized mutual information (NMI) [47], which is a measure of the mutual dependence between the two variables. It is observed that our model slightly outperforms all previous results and the error rate is very close to the ideal results mentioned in [78], which is 55/1222. This shows that PCABM can effectively incorporate the degree information into a specific pairwise covariate and provide significant improvement over the vanilla SBM, whose NMI is only 0.0001 as reported in [80].

Table 3.2: Performance comparison on political blogs data

	ARI	NMI	Errors
Karrer and Newman (2011) [80]	-	0.72	-
Zhao et al. (2012) [147]	0.819	-	-
Jin (2015) [78]	0.819	0.725	58
PCABM.MLE	**0.825**	**0.737**	**56**

3.5.2 Example 2: school friendship

For real networks, people often use certain nodal labels as the ground ‘truth’ for community labels to evaluate the performance of various community detection methods. However, there could be different ‘true’ community assignments based on different nodal labels (e.g., gender, job, and age). [115] mentioned that communities and the covariates may capture different aspects of the network, which is inline with the idea presented here. To examine whether PCABM can discover different community structures, in our second example, we treat one covariate as the indicator for the unknown ‘true’ community assignments while using the remaining covariates as the known covariates in our PCABM model.

The dataset is a friendship network of school students, which is from the National Longitudinal

Study of Adolescent to Adult Health (Add Health). It contains 795 students from a high school (Grades 9-12) and its feeder middle school (Grade 7-8). The pairwise covariates include grade, gender, ethnicity and number of friends nominated (up to 10). We focused on the largest connected component with at least one covariate non-missing and treat the network as undirected, resulting in a network with 777 nodes and 4124 edges. For the nodes without gender, we let them to be female, which is the smaller group. For those without grade, we generated a random grade according to their schools.

Different from traditional community detection algorithms which can only detect one underlying community structure, PCABM provides us with more flexibility to uncover different community structures by controlling different covariates. Our intuition is that social network is usually determined by multiple underlying structures, so it cannot be simply explained by one covariate. Sometimes one community structure seems to dominate the network, but if we control the covariate associated with this structure, we may discover other interesting community structures.

In this example, we conducted three community detection experiments. In each experiment, out of the three nodal labels (school, ethnicity and gender), one was viewed as the indicator for the underlying community, and community detection was carried out by using the pairwise covariates constructed using the other two covariates. For gender, school and ethnicity, we created indicator variables to represent whether the corresponding covariate values were the same for the pair of nodes. Besides, we considered the number of nominated friends in all experiments and grade for predicting ethnicity and gender. For number of nominated friends, we used $\log(n_i+1)+\log(n_j+1)$ as one pairwise covariate, where n_i is the number of nominated friends for the i-th student. '+1' was used here because some students didn't nominate anyone. For grades, we used the absolute difference to form a pairwise covariate. Using random initial community labels, we derived the estimates $\hat{\gamma}$ in each experiment. In Tables 3.3 and 3.4, we show respectively the estimates when school and ethnicity are taken as the targeted community.

In both tables, the standard error is calculated by Theorem 3.2, with the theoretical values replaced by the estimated counterparts. Thus, we can calculate the t value for each coefficient

Table 3.3: Inference results when school is targeted community

Covariate	Estimate	Std. Error	*t* value	Pr(> |t|)
White	1.251	0.043	29.002	< 0.001***
Black	1.999	0.051	38.886	< 0.001***
Hispanic	0.048	0.523	0.091	0.927
Others	0.019	0.543	0.035	0.972
Gender	0.192	0.034	-5.620	< 0.001***
Nomination	0.438	0.024	18.584	< 0.001***

Table 3.4: Inference results when ethnicity is targeted community

Covariate	Estimate	Std. Error	*t* value	Pr(> |t|)
School	1.005	0.076	-13.168	< 0.001***
Grade	-1.100	0.028	-39.182	< 0.001***
Gender	0.198	0.034	-5.813	< 0.001***
Nomination	0.498	0.023	21.679	< 0.001***

and perform statistical tests. We can see that in both experiments, the coefficients for gender and the number of nominations are similar. They both appear to be significant in the creation of the friendship network. The significant positive coefficient of nominations shows that students with a large number of nominations tend to be friends with each other, which is intuitive. The positive coefficients of gender and school shows students of the same gender and school are more likely to be friends with each other, which is in line with our expectations. The negative coefficient of grade means that students with closer grades are more likely to be friends with each other. If we take a closer look at the coefficients of different ethnic groups in Table 3.3, we find that only those corresponding to white and black are significant. This is understandable if we observe that among 777 students, 476 are white and 221 are black. As for school and grade, students in the same school or grade are more likely to be friends with each other, which is as we expected.

The network is divided into two communities each time (we only look at white and black students in the second experiment because the sizes for other ethnicities are very small). In Figure 3.5, school, ethnicity and gender are targeted communities respectively. We use different shades to distinguish true communities. Predicted communities are separated by the middle dash line, so the ideal split would be shades vs. tints on two sides. By this criteria, our model performs pretty well for the first two cases, but fails to distinguish different genders. Also, more edges between two communities in the third plot indicates bad performance. It can either be the existence of another unknown variables or gender's contribution to the network structure is insignificant given the covariates considered.

The results in terms of ARI are shown in Table 3.5. Note that, for all other methods, we would get only one community structure, whose performance is doomed to be bad for capturing different community structures. Also, to test the robustness of our method, in the experiment of detecting the ethnicity community, we tried to use the square of the grade difference which led to almost the same ARI.

Table 3.5: ARI comparison on school friendship data

	School	Race	Gender
PCABM.MLE	**0.909**	**0.914**	**0.030**
SBM.MLE	0.048	0.138	-0.001
SBM.SC	0.043	-0.024	0.000
DCBM.MLE	**0.909**	0.001	0.002
SCORE	0.799	0.012	0.011

3.6 Discussion

In this chapter, we extended the classical stochastic block model to let the connection probability between nodes not only depend on communities but also on the pairwise covariates. We proved consistency in terms of both coefficient estimates and community label assignments for MLE under PCABM. Also, we introduced a fast spectral method SCWA with theoretical justification, which may serve as a good initial solution for the likelihood method.

Though we assumed A_{ij}'s are non-negative integers, it can be relaxed to be any non-negative number, and the proof still holds. Also, we would like to consider unbalanced community sizes $n_{\min}/n_{\max} = o(1)$ or when the number of communities K_n diverges.

One possible future work is to extend PCABM to directed network. Besides, for the bounded degree case where $\varphi_n = O(1)$, it is of great interest to discuss the properties of estimators under

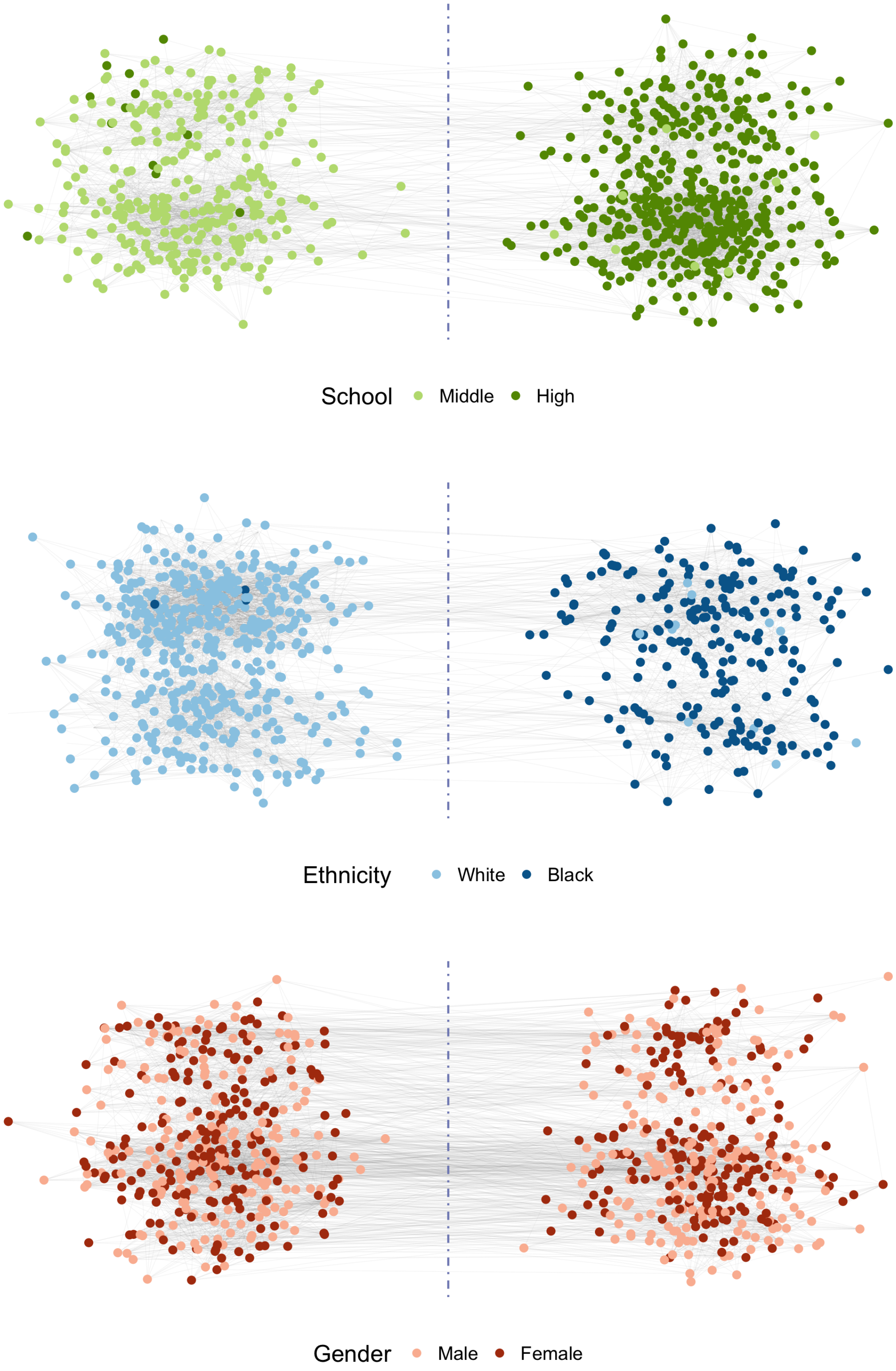


Figure 3.5: Community detection with different pairwise covariates

our framework. When we have high-dimensional pairwise covariates, adding a penalty term to perform variable selection is also worth investigating.

Another interesting issue is the choice of the number of communities K which is assumed to be known in this thesis. However, in practice, it would be desirable to have an automatic selection procedure for choosing K. Some recent efforts toward this direction include [121, 87, 134, 90, 39, 92, 143]. It would be interesting to study some of the methods under PCABM.

3.7 Proofs

This section contains proofs of Theorems 3.1, 3.2, 3.3, 3.4 and 3.5.

3.7.1 Proof of Theorem 3.1 and Theorem 3.2

Proof. In the following proof, we will use $\hat{\boldsymbol{\gamma}}$ instead of $\hat{\boldsymbol{\gamma}}(\mathbf{e})$ for simplicity. Define the empirical version of $\theta(\boldsymbol{\gamma})$, $\boldsymbol{\mu}(\boldsymbol{\gamma})$ and $\Sigma(\boldsymbol{\gamma})$ as

$$
\begin{aligned}
\hat{\theta}^{kl}(\boldsymbol{\gamma}) &= \sum_{(u,v)\in s_{\mathbf{e}}(k,l)} e^{\mathbf{z}_{uv}^T\boldsymbol{\gamma}}/|s_{\mathbf{e}}(k,l)|,\\
\hat{\boldsymbol{\mu}}^{kl}(\boldsymbol{\gamma}) &= \sum_{(u,v)\in s_{\mathbf{e}}(k,l)} \mathbf{z}_{uv}e^{\mathbf{z}_{uv}^T\boldsymbol{\gamma}}/|s_{\mathbf{e}}(k,l)|,\\
\hat{\Sigma}^{kl}(\boldsymbol{\gamma}) &= \sum_{(u,v)\in s_{\mathbf{e}}(k,l)} \mathbf{z}_{uv}\mathbf{z}_{uv}^T e^{\mathbf{z}_{uv}^T\boldsymbol{\gamma}}/|s_{\mathbf{e}}(k,l)|.
\end{aligned}
$$

For fixed $\boldsymbol{\gamma}$, under Condition 3.1 and 3.4, for $\forall k,l\in[K]$, we know the weak law of large numbers holds, i.e., $\hat{\theta}^{kl}(\boldsymbol{\gamma})\xrightarrow{p}\theta(\boldsymbol{\gamma})$, $\hat{\boldsymbol{\mu}}^{kl}(\boldsymbol{\gamma})\xrightarrow{p}\boldsymbol{\mu}(\boldsymbol{\gamma})$ and $\hat{\Sigma}^{kl}(\boldsymbol{\gamma})\xrightarrow{p}\Sigma(\boldsymbol{\gamma})$.

First, we calculate the first and second order derivative of $\ell_{\mathbf{e}}$ w.r.t. $\boldsymbol{\gamma}$ as

$$
\begin{aligned}
\ell_{\mathbf{e}}'(\boldsymbol{\gamma}) &= \frac{\partial \ell_{\mathbf{e}}}{\partial\boldsymbol{\gamma}} = \sum_{i<j} A_{ij}\mathbf{z}_{ij} - \sum_{i<j} A_{ij}\hat{\boldsymbol{\mu}}^{e_ie_j}(\boldsymbol{\gamma})/\hat{\theta}^{e_ie_j}(\boldsymbol{\gamma}),\\
\ell_{\mathbf{e}}''(\boldsymbol{\gamma}) &= \frac{\partial^2 \ell_{\mathbf{e}}}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = \sum_{kl}\frac{O_{kl}(\mathbf{e})}{2\hat{\theta}^{kl}(\boldsymbol{\gamma})^2}\left[\hat{\boldsymbol{\mu}}^{kl}(\boldsymbol{\gamma})^{\otimes 2} - \hat{\theta}^{kl}(\boldsymbol{\gamma})\hat{\Sigma}^{kl}(\boldsymbol{\gamma})\right].
\end{aligned}
$$

Then conditioned on Z, we define the expectation and variance of $\ell_{\mathbf{e}}'(\boldsymbol{\gamma})$ as $\boldsymbol{\mu}_{\ell}(\mathbf{e},\boldsymbol{\gamma}) = \mathbb{E}[\ell_{\mathbf{e}}'(\boldsymbol{\gamma})|Z]$

and $\Sigma_\ell(\mathbf{e}, \boldsymbol{\gamma}) = \mathrm{var}[\ell_{\mathbf{e}}'(\boldsymbol{\gamma})|Z]$.

$$\begin{aligned}
\frac{\mu_\ell(\mathbf{e},\boldsymbol{\gamma})}{N_n\rho_n} =&(N_n\rho_n)^{-1}\sum_{i<j} B_{c_ic_j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\left[\mathbf{z}_{ij} - \frac{\hat{\boldsymbol{\mu}}^{e_ie_j}(\boldsymbol{\gamma})}{\hat{\theta}^{e_ie_j}(\boldsymbol{\gamma})}\right]\\
=&N_n^{-1}\sum_{i<j}\Omega_{c_ic_j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\left[\mathbf{z}_{ij} - \frac{\boldsymbol{\mu}(\boldsymbol{\gamma})}{\theta(\boldsymbol{\gamma})}\right] + o(1)\\
=&N_n^{-1}\sum_{i<j}\Omega_{c_ic_j}[(e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\mathbf{z}_{ij} - \boldsymbol{\mu}(\boldsymbol{\gamma})) + \frac{\boldsymbol{\mu}(\boldsymbol{\gamma})}{\theta(\boldsymbol{\gamma})}(\theta(\boldsymbol{\gamma}) - e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0})] + o(1),\\
\frac{\Sigma_\ell(\mathbf{e},\boldsymbol{\gamma})}{N_n\rho_n} =&(N_n\rho_n)^{-1}\sum_{i<j} B_{c_ic_j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\left[\mathbf{z}_{ij} - \frac{\hat{\boldsymbol{\mu}}^{e_ie_j}(\boldsymbol{\gamma})}{\hat{\theta}^{e_ie_j}(\boldsymbol{\gamma})}\right]^{\otimes 2}\\
=&N_n^{-1}\sum_{i<j}\Omega_{c_ic_j} e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\left[\mathbf{z}_{ij} - \frac{\boldsymbol{\mu}(\boldsymbol{\gamma})}{\theta(\boldsymbol{\gamma})}\right]^{\otimes 2} + o(1)\\
=&(2N_n)^{-1}\sum_{kl}\Omega_{kl}\sum_{(i,j)\in s_{\mathbf{c}}(k,l)}[e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\mathbf{z}_{ij}^{\otimes 2} - 2e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\mathbf{z}_{ij}\boldsymbol{\mu}(\boldsymbol{\gamma})^T/\theta(\boldsymbol{\gamma})\\
&+ e^{\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0}\boldsymbol{\mu}(\boldsymbol{\gamma})^{\otimes 2}/\theta(\boldsymbol{\gamma})^2] + o(1).
\end{aligned}$$

By LLN, $\lim_{n\to\infty}\frac{\mu_\ell(\mathbf{e},\boldsymbol{\gamma})}{N_n\rho_n}$ and $\lim_{n\to\infty}\frac{\Sigma_\ell(\mathbf{e},\boldsymbol{\gamma})}{N_n\rho_n}$ exist, and we use $\boldsymbol{\mu}_\infty(\boldsymbol{\gamma})$ and $\Sigma_\infty(\boldsymbol{\gamma})$ to denote them respectively, where

$$\begin{aligned}
\boldsymbol{\mu}_\infty(\boldsymbol{\gamma}) &= \sum_{ab}\Omega_{ab}\pi_a\pi_b\left[\boldsymbol{\mu}(\boldsymbol{\gamma}^0) - \frac{\boldsymbol{\mu}(\boldsymbol{\gamma})}{\theta(\boldsymbol{\gamma})}\theta(\boldsymbol{\gamma}^0)\right],\\
\Sigma_\infty(\boldsymbol{\gamma}) &= \sum_{ab}\Omega_{ab}\pi_a\pi_b\left[\Sigma(\boldsymbol{\gamma}^0) - \frac{2\boldsymbol{\mu}(\boldsymbol{\gamma}^0)\boldsymbol{\mu}(\boldsymbol{\gamma})^T}{\theta(\boldsymbol{\gamma})} + \frac{\theta(\boldsymbol{\gamma}^0)\boldsymbol{\mu}(\boldsymbol{\gamma})^{\otimes 2}}{\theta(\boldsymbol{\gamma})^2}\right],
\end{aligned}$$

Specifically, at $\boldsymbol{\gamma}^0$, we have $\boldsymbol{\mu}_\infty(\boldsymbol{\gamma}^0) = \mathbf{0}$ and $\Sigma_\infty(\boldsymbol{\gamma}^0) = \sum_{ab}\Omega_{ab}\pi_a\pi_b[\Sigma(\boldsymbol{\gamma}^0) - \theta(\boldsymbol{\gamma}^0)^{-1}\boldsymbol{\mu}(\boldsymbol{\gamma}^0)^{\otimes 2}]$. By Condition 3.3, $\Sigma_\infty(\boldsymbol{\gamma})$ is positive definite in a neighborhood of $\boldsymbol{\gamma}^0$.

Using Taylor expansion,

$$\ell_{\mathbf{e}}'(\hat{\boldsymbol{\gamma}}) - \ell_{\mathbf{e}}'(\boldsymbol{\gamma}^0) = \ell_{\mathbf{e}}''(\bar{\boldsymbol{\gamma}})(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0),$$

where $\bar{\boldsymbol{\gamma}} = q\boldsymbol{\gamma}^0 + (1-q)\hat{\boldsymbol{\gamma}}$ for some $q \in [0,1]$. Noticing that $\ell_{\mathbf{e}}'(\hat{\boldsymbol{\gamma}}) = 0$, so

$$\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0 = \left(\frac{\ell_{\mathbf{e}}''(\bar{\boldsymbol{\gamma}})}{N_n\rho_n}\right)^{-1}\frac{\ell_{\mathbf{e}}'(\boldsymbol{\gamma}^0)}{N_n\rho_n} \xrightarrow{p} \mathbf{0}.$$

To prove asymptotic normality, we rewrite above formula as

$$\sqrt{N_n\rho_n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) = -\left[\frac{1}{N_n\rho_n}\ell''_{\mathbf{e}}(\bar{\boldsymbol{\gamma}})\right]^{-1} \times \left[\frac{1}{\sqrt{N_n\rho_n}}\ell'_{\mathbf{e}}(\boldsymbol{\gamma}^0)\right].$$

Since $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^0$, we have $N_n\rho_n\ell''^{-1}_{\mathbf{e}}(\bar{\boldsymbol{\gamma}}) \xrightarrow{p} \Sigma_2(\boldsymbol{\gamma}^0)$. Also, conditioned on Z, by C.L.T.,

$$\frac{1}{\sqrt{N_n\rho_n}}\ell'_{\mathbf{e}}(\boldsymbol{\gamma}^0) = \sqrt{N_n\rho_n}\left[\frac{1}{N_n\rho_n}\ell'_{\mathbf{e}}(\boldsymbol{\gamma}^0)\right] \xrightarrow{d} N(0, \Sigma_2(\boldsymbol{\gamma}^0)).$$

Thus,

$$\sqrt{N_n\rho_n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0) \xrightarrow{d} N(0, \Sigma_\infty^{-1}(\boldsymbol{\gamma}^0)).$$

□

3.7.2 Some concentration inequalities and notations

For the purpose of later proofs, we introduce some concentration inequalities and necessary notations first.

One inequality that we will apply repeatedly is extended version of Bernstein inequality for unbounded random variables introduced in [138].

Lemma 3.8 (Bernstein inequality). *Suppose $X_1, \cdots, X_n$ are independent random variables with $\mathbb{E}X_i = 0$ and $\mathbb{E}|X_i|^k \leq \frac{1}{2}\mathbb{E}X_i^2 L^{k-2} k!$ for $k \geq 2$. For $M \geq \sum_{i \leq n} \mathbb{E}X_i^2$ and $x \geq 0$,*

$$\Pr\left(\sum_{i\leq n} X_i \geq x\right) \leq \exp\left(-\frac{x^2}{2(M + xL)}\right).$$

To show that all Poisson distributions satisfy the above Bernstein condition uniformly under some constant $\bar{L}$, we give the following lemma.

Lemma 3.9 (Bernstein condition). *Assume $A \sim Pois(\lambda)$, let $X = A - \lambda$, then for any $0 < \lambda < 1/2$, there exists a constant $\bar{L} > 0$ s.t. for any integer $k > 2$, $\mathbb{E}[|X^k|] \leq \mathbb{E}[X^2]\bar{L}^{k-2}k!/2$.*

Proof.

$$
\begin{aligned}
&\frac{2\mathbb{E}[|A-\lambda|^k]}{\lambda k!} = \frac{2}{\lambda k!}\mathbb{E}[(A-\lambda)^k|A \geq 1]\Pr(A \geq 1) + \frac{2\lambda^{k-1}e^{-\lambda}}{k!} \\
\leq &\frac{2}{\lambda k!}\mathbb{E}[A^k|A \geq 1]\Pr(A \geq 1) + e^{-\lambda} = \frac{2}{\lambda k!}\mathbb{E}[A^k] + e^{-\lambda}\frac{2}{k!}\sum_{i=1}^{k}\left\{\begin{matrix} k \\ i \end{matrix}\right\}\lambda^{i-1} + e^{-\lambda} \\
\leq &\frac{1}{k!}\sum_{i=1}^{k}\binom{k}{i}i^{k-i}\lambda^{i-1} + e^{-\lambda} \leq \frac{e^{k-1}}{k^k}\sum_{i=1}^{k}\left(\frac{ek}{i}\right)^i i^{k-i}\lambda^{i-1} + e^{-\lambda} \\
= &\sum_{i=1}^{k} e^{i+k-1}i^{k-2i}k^{i-k}\lambda^{i-1} + e^{-\lambda} < \sum_{i=1}^{k} e^{i+k-1}e^{-i}\lambda^{i-1} + e^{-\lambda} = e^{k-1}\frac{1-\lambda^k}{1-\lambda} + e^{-\lambda} \\
\leq &\frac{e^{k-1}}{1-\lambda} + 1 \leq \left(\frac{e^2+1}{1-\lambda}\right)^{k-2}.
\end{aligned}
$$

Notice that when λ is bounded away from 1, say $\lambda < 1/2$, we can simply set $\bar{L} = 2(e^2+1)$, then Bernstein condition is satisfied uniformly for all λ. □

We introduce some notations. Let $|\mathbf{e} - \mathbf{c}| = \sum_{i=1}^{n} \mathbb{1}_{e_i \neq c_i}$. Given a community assignment $\mathbf{e} \in [K]^n$, we define $V(\mathbf{e}) \in \mathbb{R}^{K \times K}$ with its elements being

$$V_{ka}(\mathbf{e}) \equiv \frac{\sum_{i=1}^{n} \mathbb{1}_{e_i=k, c_i=a}}{\sum_{i=1}^{n} \mathbb{1}_{c_i=a}} = \frac{R_{ka}(\mathbf{e})}{\pi_a(\mathbf{c})}.$$

One can view V as the empirical conditional distribution of $\mathbf{e}$ given $\mathbf{c}$. We can see that $V(\mathbf{e}) = R(\mathbf{e})D(\mathbf{c})^{-1}$. Also, note that $V(\mathbf{e})^T\mathbf{1} = \mathbf{1}$, $V(\mathbf{e})\boldsymbol{\pi}(\mathbf{c}) = \boldsymbol{\pi}(\mathbf{e})$ and $V(\mathbf{c}) = I_K$. For the convenience of later proof, we also define $W(\mathbf{c}) = D(\mathbf{c})\Omega D(\mathbf{c})$ and

$$
\begin{aligned}
\hat{T}(\mathbf{e}) \equiv &R(\mathbf{e})\Omega R(\mathbf{e})^T = V(\mathbf{e})W(\mathbf{c})V(\mathbf{e})^T, \\
\hat{S}(\mathbf{e}) \equiv &V(\mathbf{e})\boldsymbol{\pi}(\mathbf{c})\boldsymbol{\pi}(\mathbf{c})^T V(\mathbf{e})^T.
\end{aligned}
$$

Replacing the empirical distribution $\boldsymbol{\pi}(\mathbf{c})$ by the true distribution $\boldsymbol{\pi}$, we define $W_0 = D(\boldsymbol{\pi})\Omega D(\boldsymbol{\pi})$, and $T(\mathbf{e}), S(\mathbf{e}) \in \mathbb{R}^{K\times K}$ as

$$\begin{aligned} T(\mathbf{e}) \equiv& V(\mathbf{e}) W_0 V(\mathbf{e})^T, \\ S(\mathbf{e}) \equiv& V(\mathbf{e}) \boldsymbol{\pi}\boldsymbol{\pi}^T V(\mathbf{e})^T. \end{aligned}$$

The population version of $F\left(\frac{O}{2N_n\rho_n}, \frac{E}{2N_n}\right)$ is $F(\theta(\boldsymbol{\gamma}^0)T(\mathbf{e}), \theta(\hat{\boldsymbol{\gamma}})S(\mathbf{e}))$. To measure the discrepancy between empirical and population version of F, we define $X(\mathbf{e}), Y(\mathbf{e}, \hat{\boldsymbol{\gamma}}) \in \mathbb{R}^{K\times K}$ to be the rescaled difference between O, E and their expectations

$$\begin{aligned} X(\mathbf{e}) \equiv& \frac{O(\mathbf{e})}{2N_n\rho_n} - \theta(\boldsymbol{\gamma}^0)\hat{T}(\mathbf{e}), \\ Y(\mathbf{e}, \hat{\boldsymbol{\gamma}}) \equiv& \frac{E(\mathbf{e}, \hat{\boldsymbol{\gamma}})}{2N_n} - \theta(\hat{\boldsymbol{\gamma}})\hat{S}(\mathbf{e}). \end{aligned}$$

Before we establish bound for $Y(\mathbf{e}, \hat{\boldsymbol{\gamma}})$, we need to consider the bound for $\hat{\boldsymbol{\gamma}}$. The following is a direct corollary of Theorem 3.2. Conditioned on $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\|_\infty \leq \phi$, we have $|e^{\mathbf{z}_{ij}\hat{\boldsymbol{\gamma}}} - \mathbb{E}e^{\mathbf{z}_{ij}\hat{\boldsymbol{\gamma}}}| \leq \exp\{p\alpha(\phi + \|\boldsymbol{\gamma}^0\|_\infty)\} \equiv \chi$ uniformly for any $i, j \in [n]$ and $\hat{\boldsymbol{\gamma}}$. Under this condition, we establish Lemma 3.11 using Bernstein inequality.

Lemma 3.10. *For any constant $\phi > 0$, $\exists$ positive constants C_ϕ and v_ϕ s.t.,* $\Pr(\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^0\|_\infty > \phi) < C_\phi \exp(-v_\phi N_n\rho_n)$.

Lemma 3.11.

$$\Pr(\max_{\mathbf{e}} \|X(\mathbf{e})\|_\infty \geq \epsilon) \leq 2K^{n+2}\exp(-C_1\epsilon^2 N_n\rho_n) \tag{3.6}$$

for $\epsilon < \beta_u\|\Omega\|_{\max}/\bar{L}$.

$$\Pr(\max_{|\mathbf{e}-\mathbf{c}|\leq m} \|X(\mathbf{e}) - X(\mathbf{c})\|_\infty \geq \epsilon) \leq 2\binom{n}{m}K^{m+2}\exp\left(-\frac{C_3 n}{m}\epsilon^2 N_n\rho_n\right) \tag{3.7}$$

for $\epsilon < \eta m/n$, where $\eta = 2\beta_u \|\Omega\|_{\max}/\bar{L}$.

$$\Pr(\max_{|\mathbf{e}-\mathbf{c}|\le m} \|X(\mathbf{e}) - X(\mathbf{c})\|_\infty \ge \epsilon) \le 2\binom{n}{m} K^{m+2} \exp\left(-C_4 \epsilon N_n \rho_n\right) \tag{3.8}$$

for $\epsilon \ge \eta m/n$.

$$\Pr(\max_{\mathbf{e}} \|Y(\mathbf{e}, \hat{\boldsymbol{\gamma}})\|_\infty \ge \epsilon) \le 2K^{n+2} \exp(-C_2 \epsilon^2 N_n \rho_n) \tag{3.9}$$

for $\epsilon < \chi\kappa_2^2$.

$$\Pr(\max_{|\mathbf{e}-\mathbf{c}|\le m} \|Y(\mathbf{e}, \hat{\boldsymbol{\gamma}}) - Y(\mathbf{c}, \hat{\boldsymbol{\gamma}})\|_\infty \ge \epsilon) \le 2\binom{n}{m} K^{m+2} \exp\left(-\frac{C_5 n}{m} \epsilon^2 N_n \rho_n\right) \tag{3.10}$$

for $\epsilon < \frac{2\chi m}{n}$.

$$\Pr(\max_{|\mathbf{e}-\mathbf{c}|\le m} \|Y(\mathbf{e}, \hat{\boldsymbol{\gamma}}) - Y(\mathbf{c}, \hat{\boldsymbol{\gamma}})\|_\infty \ge \epsilon) \le 2\binom{n}{m} K^{m+2} \exp\left(-C_6 \epsilon N_n \rho_n\right) \tag{3.11}$$

for $\epsilon \ge \frac{2\chi m}{n}$.

Proof. The proofs are all given conditioned on $|e^{\mathbf{z}_{ij}\hat{\gamma}} - \mathbb{E}e^{\mathbf{z}_{ij}\hat{\gamma}}| \le \chi$. By combining Lemma 3.10, we could have the conclusion directly. For any $\hat{\boldsymbol{\gamma}}$, by Bernstein inequality, when $\epsilon < \chi\kappa_2^2$,

$$\begin{aligned}\Pr(|Y_{kl}(\mathbf{e}, \hat{\boldsymbol{\gamma}})| \ge \epsilon) \le & 2\exp\left(-\frac{\frac{1}{2}(2N_n\epsilon)^2}{|s_{\mathbf{e}}(k,l)|\chi^2 + \frac{2}{3}\chi N_n \epsilon}\right) \\ \le & 2\exp\left(-\frac{6N_n\epsilon^2}{3\kappa_2^2\chi^2 + 2\chi\epsilon}\right) \le 2\exp\left(-\frac{6}{5\kappa_2^2\chi^2}\epsilon^2 N_n\right).\end{aligned}$$

Let $X^1(\mathbf{e}) = \frac{O(\mathbf{e}) - \mathbb{E}[O(\mathbf{e})|Z]}{2N_n\rho_n}$ and $X^2(\mathbf{e}) = X(\mathbf{e}) - X^1(\mathbf{e})$, and we establish bound for $X^1(\mathbf{e})$ and $X^2(\mathbf{e})$ respectively. By Lemma 3.8, for any $k, l \in [K]$, let $M = \beta_u \|\Omega\|_{\max} |s_{\mathbf{e}}(k,l)|\rho_n$, $L = \bar{L}$, and $x = 2N_n\rho_n\epsilon$, then for $\epsilon < \beta_u\|\Omega\|_{\max}/\bar{L}$,

$$\Pr(X^1_{kl}(\mathbf{e}) \ge \epsilon) \le \exp\left(-\frac{4N_n^2\rho_n^2\epsilon^2}{2(\beta_u\|\Omega\|_{\max}|s_{\mathbf{e}}(k,l)|\rho_n + 2N_n\rho_n\epsilon\bar{L})}\right)$$

$$\leq \exp\left(-\frac{N_n \rho_n \epsilon^2}{\beta_u \|\Omega\|_{\max} + \epsilon \bar{L}}\right) \leq \exp\left(-\frac{\epsilon^2 N_n \rho_n}{2\beta_u \|\Omega\|_{\max}}\right).$$

Notice that $|X_{kl}^2(\mathbf{e})|/\|\Omega\|_{\max} \leq |Y_{kl}(\mathbf{e}, \boldsymbol{\gamma}^0)|$. Thus, for $\epsilon < \chi\kappa_2^2\|\Omega\|_{\max}$,

$$\Pr(|X_{kl}^2(\mathbf{e})| \geq \epsilon) \leq \Pr\left(|Y_{kl}(\mathbf{e}, \boldsymbol{\gamma}^0)| \geq \frac{\epsilon}{\|\Omega\|_{\max}}\right) \leq 2\exp\left(-\frac{6}{5\kappa_2^2\chi^2\|\Omega\|_{\max}}\epsilon^2 N_n\right).$$

Thus, the bound of $X(\mathbf{e})$ will be dominated by $X^1(\mathbf{e})$, and we will ignore the second term in the bound because it is just a small order and can be absorbed into the first one.

Similar to the arguments in [147], for $|\mathbf{e} - \mathbf{c}| \leq m$, $\epsilon < \frac{2\chi m}{n}$,

$$\Pr(|Y_{kl}(\mathbf{e}, \hat{\boldsymbol{\gamma}}) - Y_{kl}(\mathbf{c}, \hat{\boldsymbol{\gamma}})| \geq \epsilon) \leq 2\exp\left(-\frac{6N_n\epsilon^2}{6\chi^2 mn/N_n + 2\chi\epsilon}\right)$$
$$\leq 2\exp\left(-\frac{3(n-1)}{8\chi^2 m}\epsilon^2 N_n\right) \leq 2\exp\left(-\frac{n}{4\chi^2 m}\epsilon^2 N_n\right).$$

For $\epsilon \geq \frac{2\chi m}{n}$,

$$\Pr(|Y_{kl}(\mathbf{e}, \hat{\boldsymbol{\gamma}}) - Y_{kl}(\mathbf{c}, \hat{\boldsymbol{\gamma}})| \geq \epsilon) \leq 2\exp\left(-\frac{6N_n\epsilon^2}{6\chi^2 mn/N_n + 2\chi\epsilon}\right) \leq 2\exp\left(-\frac{3}{4\chi}\epsilon N_n\right).$$

Also, for $\epsilon < \frac{2\beta_u\|\Omega\|_{\max} m}{n\bar{L}}$,

$$\Pr(|X_{kl}^1(\mathbf{e}) - X_{kl}^1(\mathbf{c})| \geq \epsilon) \leq \exp\left(-\frac{N_n\rho_n\epsilon^2}{\beta_u\|\Omega\|_{\max} mn/N_n + \epsilon\bar{L}}\right)$$
$$\leq \exp\left(-\frac{n-1}{2\beta_u\|\Omega\|_{\max} m}\epsilon^2 N_n\rho_n\right) \leq \exp\left(-\frac{n}{4\beta_u\|\Omega\|_{\max} m}\epsilon^2 N_n\rho_n\right).$$

For $\epsilon \geq \frac{2\beta_u\|\Omega\|_{\max} m}{n\bar{L}}$,

$$\Pr(|X_{kl}^1(\mathbf{e}) - X_{kl}^1(\mathbf{c})| \geq \epsilon) \leq \exp\left(-\frac{N_n\rho_n\epsilon^2}{\beta_u\|\Omega\|_{\max} mn/N_n + \epsilon\bar{L}}\right) \leq \exp\left(-\frac{1}{3\bar{L}}\epsilon N_n\rho_n\right).$$

We will omit the bound for $|X_{kl}^2(\mathbf{e}) - X_{kl}^2(\mathbf{c})|$ since it’s a smaller order.

□

3.7.3 Proof of Theorem 3.3

Proof. We divide the proof into three steps.

Step 1 : sample and population version comparison. We prove $\exists\ \epsilon_n \to 0$, such that

$$\mathrm{P}\left(\max_{\mathbf{e}}\left|F\left(\frac{O(\mathbf{e})}{2N_n\rho_n},\frac{E(\mathbf{e},\hat{\boldsymbol{\gamma}})}{2N_n}\right) - F(\theta(\boldsymbol{\gamma}^0)T(\mathbf{e}),\theta(\boldsymbol{\gamma}^0)S(\mathbf{e}))\right| \le \epsilon_n\right) \to 1, \tag{3.12}$$

if $\varphi_n \to \infty$ and $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}^0$.

Since

$$\begin{aligned}
&\left|F\left(\frac{O(\mathbf{e})}{2N_n\rho_n},\frac{E(\mathbf{e},\hat{\gamma})}{2N_n}\right) - \theta(\boldsymbol{\gamma}^0)F(T(\mathbf{e}),S(\mathbf{e}))\right| \\
\le &\left|F\left(\frac{O(\mathbf{e})}{2N_n\rho_n},\frac{E(\mathbf{e},\hat{\gamma})}{2N_n}\right) - F(\theta(\boldsymbol{\gamma}^0)\hat{T}(\mathbf{e}),\theta(\hat{\gamma})\hat{S}(\mathbf{e}))\right| \\
&+\left|F(\theta(\boldsymbol{\gamma}^0)\hat{T}(\mathbf{e}),\theta(\hat{\gamma})\hat{S}(\mathbf{e})) - \theta(\boldsymbol{\gamma}^0)F(\hat{T}(\mathbf{e}),\hat{S}(\mathbf{e}))\right| \\
&+\theta(\boldsymbol{\gamma}^0)\left|F(\hat{T}(\mathbf{e}),\hat{S}(\mathbf{e})) - F(T(\mathbf{e}),S(\mathbf{e}))\right|,
\end{aligned}$$

it is sufficient to bound these three terms uniformly. By Lipschitz continuity,

$$\left|F\left(\frac{O(\mathbf{e})}{2N_n\rho_n},\frac{E(\mathbf{e},\hat{\gamma})}{2N_n}\right) - \theta(\boldsymbol{\gamma}^0)F\left(\hat{T}(\mathbf{e}),\hat{S}(\mathbf{e})\right)\right| \le M_1\|X(\mathbf{e})\|_\infty + M_2\|Y(\mathbf{e},\hat{\gamma})\|_\infty, \tag{3.13}$$

$$\left|F(\theta(\boldsymbol{\gamma}^0)\hat{T}(\mathbf{e}),\theta(\hat{\gamma})\hat{S}(\mathbf{e})) - \theta(\boldsymbol{\gamma}^0)F(\hat{T}(\mathbf{e}),\hat{S}(\mathbf{e}))\right| \le M_2|\theta(\hat{\gamma}) - \theta(\boldsymbol{\gamma}^0)|\|\hat{S}(\mathbf{e})\|_\infty. \tag{3.14}$$

By (3.6) and (3.9), (3.13) converges to 0 uniformly if $\varphi_n \to \infty$. Since $\|\hat{S}(\mathbf{e})\|_\infty$ is uniformly bounded by 1, (3.14) also converges to 0 uniformly.

$$\left|F\left(\hat{T}(\mathbf{e}),\hat{S}(\mathbf{e})\right) - F\left(T(\mathbf{e}),S(\mathbf{e})\right)\right| \le M_1\|\hat{T}(\mathbf{e}) - T(\mathbf{e})\|_\infty + M_2\|\hat{S}(\mathbf{e}) - S(\mathbf{e})\|_\infty. \tag{3.15}$$

Since $\boldsymbol{\pi}(\mathbf{c}) \xrightarrow{p} \boldsymbol{\pi}$, (3.15) converges to 0 uniformly. So we prove (3.12).

Step 2 : proof of weak consistency. We prove that there exists $\delta_n \to 0$, such that

$$\Pr\left(\max_{\{\mathbf{e}:\|V(\mathbf{e})-I_K\|_1\geq\delta_n\}} F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e},\hat{\boldsymbol{\gamma}})}{2N_n}\right) < F\left(\frac{O(\mathbf{c})}{2N_n\rho_n}, \frac{E(\mathbf{c},\hat{\boldsymbol{\gamma}})}{2N_n}\right)\right) \to 1. \tag{3.16}$$

By continuity property of F and Condition 3.5, there exists $\delta_n \to 0$, such that

$$\theta(\boldsymbol{\gamma}^0)F(T(\mathbf{c}), S(\mathbf{c})) - \theta(\boldsymbol{\gamma}^0)F(T(\mathbf{e}), S(\mathbf{e})) > 2\epsilon_n$$

if $\|V(\mathbf{e}) - I_K\|_1 \geq \delta_n$, where $I_K = V(\mathbf{c})$. Thus, following (3.12),

$$\begin{aligned}
&\Pr\left(\max_{\{\mathbf{e}:\|V(\mathbf{e})-I_K\|_1\geq\delta_n\}} F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e},\hat{\boldsymbol{\gamma}})}{2N_n}\right) < F\left(\frac{O(\mathbf{c})}{2N_n\rho_n}, \frac{E(\mathbf{c},\hat{\boldsymbol{\gamma}})}{2N_n}\right)\right)\\
\geq&\Pr\left(\left|\max_{\mathbf{e}:\|V(\mathbf{e})-I_K\|_1\geq\delta_n} \theta(\boldsymbol{\gamma}^0)F(T(\mathbf{e}), S(\mathbf{e})) - \max_{\mathbf{e}:\|V(\mathbf{e})-I_K\|_1\geq\delta_n} F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e},\hat{\boldsymbol{\gamma}})}{2N_n}\right)\right| \leq \epsilon_n,\right.\\
&\left.\left|\theta(\boldsymbol{\gamma}^0)F(T(\mathbf{c}), S(\mathbf{c})) - F\left(\frac{O(\mathbf{c})}{2N_n\rho_n}, \frac{E(\mathbf{c},\hat{\boldsymbol{\gamma}})}{2N_n}\right)\right| \leq \epsilon_n\right) \to 1.
\end{aligned}$$

(3.16) implies $\Pr(\|V(\mathbf{e}) - I_K\| < \delta_n) \to 1$. Since

$$\frac{1}{n}|\mathbf{e} - \mathbf{c}| = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{c_i \neq e_i} = \sum_k \pi_k(1 - V_{kk}(\mathbf{e})) \leq \sum_k (1 - V_{kk}(\mathbf{e})) = \|V(\mathbf{e}) - I_K\|_1/2,$$

weak consistency follows.

Step 3 : proof of strong consistency.

To prove strong consistency, we need to show

$$\Pr\left(\max_{\{\mathbf{e}:0<\|V(\mathbf{e})-I_K\|_1<\delta_n\}} F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e},\hat{\boldsymbol{\gamma}})}{2N_n}\right) < F\left(\frac{O(\mathbf{c})}{2N_n\rho_n}, \frac{E(\mathbf{c},\hat{\boldsymbol{\gamma}})}{2N_n}\right)\right) \to 1. \tag{3.17}$$

Combining (3.16) and (3.17), we have

$$\Pr\left(\max_{\{\mathbf{e}:\mathbf{e}\neq\mathbf{c}\}} F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e},\hat{\boldsymbol{\gamma}})}{2N_n}\right) < F\left(\frac{O(\mathbf{c})}{2N_n\rho_n}, \frac{E(\mathbf{c},\hat{\boldsymbol{\gamma}})}{2N_n}\right)\right) \to 1,$$

which implies strong consistency.

By Lipschitz continuity and the continuity of derivative of F w.r.t. $V(\mathbf{e})$ in the neighborhood of I_K, we have

$$\begin{aligned} & F\left(\frac{O(\mathbf{e})}{2N_n\rho_n}, \frac{E(\mathbf{e}, \hat{\boldsymbol{\gamma}})}{2N_n}\right) - F\left(\frac{O(\mathbf{c})}{2N_n\rho_n}, \frac{E(\mathbf{c}, \hat{\boldsymbol{\gamma}})}{2N_n}\right) \\ =& \theta(\boldsymbol{\gamma}^0)F(\hat{T}(\mathbf{e}), \hat{S}(\mathbf{e})) - \theta(\boldsymbol{\gamma}^0)F(\hat{T}(\mathbf{c}), \hat{S}(\mathbf{c})) + \Delta(\mathbf{e}, \mathbf{c}), \end{aligned} \tag{3.18}$$

where $|\Delta(\mathbf{e}, \mathbf{c})| \leq M_3(\|X(\mathbf{e}) - X(\mathbf{c})\|_\infty) + M_4\|Y(\mathbf{e}, \hat{\boldsymbol{\gamma}}) - Y(\mathbf{c}, \hat{\boldsymbol{\gamma}})\|_\infty)$, and

$$F(T(\mathbf{e}), S(\mathbf{e})) - F(T(\mathbf{c}), S(\mathbf{c})) \leq -C\|V(\mathbf{e}) - I\|_1 + o(\|V(\mathbf{e}) - I_K\|_1).$$

Since the derivative of F is continuous w.r.t. $V(\mathbf{e})$ in the neighborhood of I_K, there exists a δ' such that,

$$F(\hat{T}(\mathbf{e}), \hat{S}(\mathbf{e})) - F(\hat{T}(\mathbf{c}), \hat{S}(\mathbf{c})) \leq -(C/2)\|V(\mathbf{e}) - I\|_1 + o(\|V(\mathbf{e}) - I_K\|_1) \tag{3.19}$$

holds when $\|\boldsymbol{\pi}(\mathbf{c}) - \boldsymbol{\pi}\|_\infty \leq \delta'$. Since $\boldsymbol{\pi}(\mathbf{c}) \to \boldsymbol{\pi}$, (3.19) holds with probability approaching 1. Combining (3.18) and (3.19), it is easy to see strong consistency follows if we can show

$$\Pr\left(\max_{\{\mathbf{e} \neq \mathbf{c}\}} |\Delta(\mathbf{e}, \mathbf{c})| \leq C\|V(\mathbf{e}) - I_K\|_1/4\right) \to 1.$$

Note $\frac{1}{n}|\mathbf{e} - \mathbf{c}| \leq \frac{1}{2}\|V(\mathbf{e}) - I_K\|_1$. So for each $m \geq 1$,

$$\begin{aligned} & \Pr\left(\max_{|\mathbf{e}-\mathbf{c}|=m} |\Delta(\mathbf{e}, \mathbf{c})| > C\|V(\mathbf{e} - I_K)\|_1/4\right) \\ \leq & \Pr\left(\max_{|\mathbf{e}-\mathbf{c}| \leq m} \|X(\mathbf{e}) - X(\mathbf{c})\|_\infty > \frac{Cm}{4M_3 n}\right) (\equiv I_1) \\ & + \Pr\left(\max_{|\mathbf{e}-\mathbf{c}| \leq m} \|Y(\mathbf{e}, \hat{\boldsymbol{\gamma}}) - Y(\mathbf{c}, \hat{\boldsymbol{\gamma}})\|_\infty > \frac{Cm}{4M_4 n}\right) (\equiv I_2). \end{aligned} \tag{3.20}$$

Let $\zeta_1 = C/4M_3$, if $\zeta_1 < \eta$, by (3.7),

$$I_1 \leq 2K^{m+2} n^m \exp\left(-\eta_1^2 \frac{C_3 m}{n} N_n \rho_n\right) = 2K^2 [K \exp(\log n - \eta_1^2 C_3 N_n \rho_n / n)]^m.$$

If $\zeta_1 > \eta$, by (3.8),

$$I_1 \leq 2K^{m+2} n^m \exp\left(-\eta_1 \frac{C_4 m}{n} N_n \rho_n\right) = 2K^2 [K \exp(\log n - \eta_1 C_4 N_n \rho_n / n)]^m.$$

Similar arguments hold for I_2 by using (3.10) and (3.11). In all cases, since $\varphi_n / \log n \to \infty$,

$$\Pr\left(\max_{\{\mathbf{e} \neq \mathbf{c}\}} |\Delta(\mathbf{e}, \mathbf{c})| > C \|V(\mathbf{e}) - I_K\|_1 / 4\right) = \sum_{m=1}^{\infty} \Pr\left(\max_{|\mathbf{e}-\mathbf{c}|=m} |\Delta(\mathbf{e}, \mathbf{c})| > C \|V(\mathbf{e}) - I_K\|_1 / 4\right) \to 0.$$

as $n \to \infty$. The proof is completed.

□

3.7.4 Proof of Theorem 3.4

By scaling $\ell_{\hat{\gamma}}(\mathbf{e})$, we have

$$\frac{1}{N_n} \ell_{\hat{\gamma}}(\mathbf{e}) = \rho_n F\left(\frac{O}{2N_n \rho_n}, \frac{E}{2N_n}\right) + (\rho_n \log \rho_n) \sum_{kl} \frac{O_{kl}(\mathbf{e})}{2N_n \rho_n} + O(n^{-1}),$$

where $F(X, Y) = \sum_{kl} X_{kl} \log\left(\frac{X_{kl}}{Y_{kl}}\right)$, $X, Y \in \mathbb{R}^{K \times K}$. Note that F is closely related to the likelihood criterion used in [147]. In our case,

$$\begin{aligned}
&F(\theta(\boldsymbol{\gamma}^0) R \Omega R^T, \theta(\boldsymbol{\gamma}^0) \boldsymbol{\pi} \boldsymbol{\pi}^T) - F(\theta(\boldsymbol{\gamma}^0) R \Omega R^T, \theta(\hat{\boldsymbol{\gamma}}) \boldsymbol{\pi} \boldsymbol{\pi}^T) \\
=& \theta(\boldsymbol{\gamma}^0) \log \frac{\theta(\boldsymbol{\gamma}^0)}{\theta(\hat{\boldsymbol{\gamma}})} (\mathbf{1}^T R \Omega R^T \mathbf{1}),
\end{aligned}$$

which is basically the population degree up to a constant. This fact shows that the consistency of $\hat{\boldsymbol{\gamma}}$ is unnecessary to ensure the consistency of community detection. Thus, we can plug in any random fixed $\hat{\boldsymbol{\gamma}}$, which is different from our general theorem. For simplicity, just assume we use the true value $\boldsymbol{\gamma}^0$ here. Observe that

$$F(\theta(\boldsymbol{\gamma}^0)R\Omega R^T, \theta(\boldsymbol{\gamma}^0)\boldsymbol{\pi}\boldsymbol{\pi}^T) = \theta(\boldsymbol{\gamma}^0)F(R\Omega R^T, \boldsymbol{\pi}\boldsymbol{\pi}^T),$$

then the form of F is exactly the same as F defined in [24], which automatically satisfies all conditions for F.

3.7.5 Proof of Theorem 3.5

This proof is adapted from [91]. First, as a direct corollary of Theorem 3.2, we have the following estimation error bound for $\hat{\boldsymbol{\gamma}}$.

Lemma 3.12. *For any constant $\eta > 0$, $\exists$ positive constants C_η and v_η s.t.,* $\Pr(\sqrt{n\rho_n}\|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty > \eta) < C_\eta \exp(-v_\eta n)$*.*

The following notations will be used in the proof.

- $W = A' - P$ and denote by w_{ij} the (i, j)-th entry of W.
- For $t > 0$, let $\mathcal{S}_t = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq t\}$ be the Euclidean ball of radius t and set $\mathcal{S} = \mathcal{S}_1$.

The proof of Theorem 3.5 is adapted from [91], so we will skip the common part and only clarify the modifications. The main idea is to bound

$$\sup_{\mathbf{x}\in\mathcal{S}} |\mathbf{x}^T(A' - P)\mathbf{x}|. \tag{3.21}$$

The original proof consists of three steps: discretization, bounding the light pairs and bounding the heavy pairs. Discretization is to reduce 3.21 to the problem of bounding the supremum of $\mathbf{x}^T(A' - P)\mathbf{y}$ for $\mathbf{x}, \mathbf{y}$ in a finite set of grid points in $\mathcal{S}$. Then we divide $\mathbf{x}, \mathbf{y}$ into light and heavy pairs and bound them respectively. We will focus on the last two steps since the first step is the same as in [91].

A. Discretization

For fixed $\delta_n \in (0,1)$, define

$$\mathcal{T} = \{\mathbf{x} = (x_1, \cdots, x_n) \in \mathcal{S} : \sqrt{n} x_i/\delta_n \in \mathbb{Z}, \forall i\},$$

where $\mathbb{Z}$ stands for the set of integers. The following lemma is the same as Lemma B.1 in [91] and we will skip the proof.

Lemma 3.13. *$\mathcal{S}_{1-\delta_n} \subset convhull(\mathcal{T})$. As a consequence, for all $W \in \mathbb{R}^{n\times n}$,*

$$\|W\| \le (1-\delta_n)^{-2} \sup_{\mathbf{x},\mathbf{y}\in\mathcal{T}} |\mathbf{x}^T W \mathbf{y}|.$$

For any $\mathbf{x}, \mathbf{y} \in \mathcal{T}$, we have

$$\mathbf{x}^T(A' - P)\mathbf{y} = \sum_{1\le i,j\le n} x_i y_j (A'_{ij} - P_{ij}).$$

We only need bound the above quantity now. We divide (x_i, y_j) into <u>light pairs</u> $\mathbb{L} = \{(i,j) : |x_i y_j| \le \sqrt{\varphi_n}/n\}$ and <u>heavy pairs</u> $\mathbb{H} = \{(i,j) : |x_i y_j| > \sqrt{\varphi_n}/n\}$. We will show that the tail for light pairs can be bounded exponentially while heavy pairs have a heavier tail. Thus, the rate of the latter one dominates.

B. Bounding the light pairs

Lemma 3.14. *Under estimation error condition, for $c > 0$, there exist constants $C_c, \nu_c > 0$ s.t.*

$$P\left(\sup_{x,y\in\mathcal{T}} \left| \sum_{(i,j)\in\mathcal{L}(x,y)} x_i y_j w_{ij} \right| \ge c\sqrt{\varphi_n}\right) \le C_c \exp\left[-\left(\nu_c - \log\left(\frac{7}{\delta}\right)\right) n\right].$$

Proof. Define $w'_{ij} = A_{ij} e^{-\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0} - P_{ij}$ and $\delta_{ij} = A_{ij} e^{-\mathbf{z}_{ij}^T \hat{\boldsymbol{\gamma}}} - A_{ij} e^{-\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}$, then $w_{ij} = w'_{ij} + \delta_{ij}$, and notice that

$$\Pr(\sum w_{ij} > 2t) \leq \Pr(\sum w'_{ij} > t) + \Pr(\sum \delta_{ij} > t),$$

so we could bound two parts respectively. Also, we need to keep in mind that

$$\sum_{i<j} u_{ij}^2 \leq \sum_{i<j} 2(x_i^2 y_j^2 + x_j^2 y_i^2) \leq 2 \sum_{1\leq i,j\leq n} x_i^2 y_j^2 = 2\|x\|_2^2 \|y\|_2^2 \leq 2.$$

<u>Step1</u> : Bound w'_{ij}.

Since

$$\begin{aligned}\sum_{i<j} \mathbb{E}[(w'_{ij} u_{ij})^2 | \mathbf{z}_{ij}] = \sum_{i<j} u_{ij}^2 \lambda_{ij} e^{-2\mathbf{z}_{ij}^T \gamma^0} &= \sum_{i<j} u_{ij}^2 P_{ij} e^{-\mathbf{z}_{ij}^T \gamma^0} \\ &\leq \rho_n \beta_l^{-1} \|\Omega\|_{\max} \sum_{i<j} u_{ij}^2 \leq 2\beta_l^{-1} \|\Omega\|_{\max} \rho_n,\end{aligned}$$

define $M = 2\beta_l^{-1}\|\Omega\|_{\max}\rho_n$, $L = 2\bar{L}\sqrt{\varphi_n}(n\beta_l)^{-1}$ and $x = c\sqrt{\varphi_n}$, we could applying Lemma 3.8 to $u_{ij} w'_{ij}$ to get

$$\Pr\left(\sum_{i<j} w'_{ij} u_{ij} \geq c\sqrt{\varphi_n}\right) \leq \exp\left(-\frac{c^2\varphi_n}{4c\bar{L}\varphi_n(n\beta_l)^{-1} + 4\rho_n\beta_l^{-1}\|\Omega\|_{\max}}\right) = \exp\left(-\frac{c^2\beta_l n}{4c\bar{L} + 4\|\Omega\|_{\max}}\right)$$

<u>Step2</u> : Bound δ_{ij}.

We consider two cases $\|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty > \eta/\sqrt{n\rho_n}$ and $\|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty \leq \eta/\sqrt{n\rho_n}$ separately. Conditioned on the second case, by choosing $\eta < (p\alpha)^{-1}$, we have

$$\begin{aligned}u_{ij}\delta_{ij} =& u_{ij}[A_{ij} e^{-\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0}(e^{\mathbf{z}_{ij}^T(\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}})} - 1)] < 2u_{ij}[A_{ij} e^{-\mathbf{z}_{ij}^T \boldsymbol{\gamma}^0} \mathbf{z}_{ij}^T(\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}})] < 2\alpha\eta p u_{ij} A_{ij}/(\beta_l\sqrt{n\rho_n}) \\ <& 2u_{ij} A_{ij}/(\beta_l\sqrt{n\rho_n})\end{aligned}$$

The first inequality is due to $|e^t - 1| < 2|t|$ when $|t| < 1$. Define $M = 2\beta_u\|\Omega\|_{\max}\rho_n \geq \sum_{i<j} u_{ij}^2 \lambda_{ij} =$

$\mathrm{var}(\sum_{i<j} u_{ij}A_{ij}|Z)$, $L = 2\bar{L}\sqrt{\varphi_n}/n$ and $x = c\sqrt{\varphi_n n\rho_n}$, by Lemma 3.8,

$$\Pr\left(\sum_{i<j} u_{ij}(A_{ij} - P_{ij}) > c\sqrt{\varphi_n n\rho_n}\right) \leq \exp\left(-\frac{c^2\varphi_n n\rho_n}{4\|\Omega\|_{\max}\beta_u\rho_n + 4c\sqrt{n\rho_n}\varphi_n\bar{L}/n}\right)$$
$$= \exp\left(-\frac{c^2 n\sqrt{n\rho_n}}{4\|\Omega\|_{\max}\beta_u/\sqrt{n\rho_n} + 4c\bar{L}}\right) \leq \exp\left(-\frac{c^2 n\sqrt{n\rho_n}}{4\|\Omega\|_{\max}\beta_u + 4c\bar{L}}\right)$$

Because

$$\sum_{i<j} u_{ij}P_{ij} \leq \rho_n\|\Omega\|_{\max}\beta_u \sum_{i<j} u_{ij} \leq \sqrt{2N_n}\rho_n\|\Omega\|_{\max}\beta_u \leq n\rho_n\|\Omega\|_{\max}\beta_u,$$

we have

$$\exp\left(-\frac{c^2 n\sqrt{n\rho_n}}{4\|\Omega\|_{\max}\beta_u + 4c\bar{L}}\right) \geq \Pr\left(\sum_{i<j} u_{ij}(A_{ij} - P_{ij}) > c\sqrt{\varphi_n n\rho_n}\right)$$
$$\geq \Pr\left(\sum_{i<j} u_{ij}A_{ij} > (c + \|\Omega\|_{\max}\beta_u)\varphi_n\right),$$

which is equivalent to $\Pr(\sum_{i<j} u_{ij}A_{ij} > c\varphi_n) \leq \exp(-C_c n\sqrt{\varphi_n})$, where C_c is constant.

Thus, for $\eta < (p\alpha)^{-1}$,

$$\Pr\left(\sum_{i<j} \delta_{ij}u_{ij} > c\sqrt{\varphi_n}\right)$$
$$= \Pr\left(\sum_{i<j} \delta_{ij}u_{ij} > c\sqrt{\varphi_n} \mid \|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty \leq \frac{\eta}{\sqrt{n\rho_n}}\right) \Pr\left(\|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty \leq \frac{\eta}{\sqrt{n\rho_n}}\right)$$
$$+ \Pr\left(\sum_{i<j} \delta_{ij}u_{ij} > c\sqrt{\varphi_n} \mid \|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty > \frac{\eta}{\sqrt{n\rho_n}}\right) \Pr\left(\|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty > \frac{\eta}{\sqrt{n\rho_n}}\right)$$
$$\leq \Pr\left(\sum_{i<j} \delta_{ij}u_{ij} > c\sqrt{\varphi_n} \mid \|\boldsymbol{\gamma}^0 - \hat{\boldsymbol{\gamma}}\|_\infty \leq \frac{\eta}{\sqrt{n\rho_n}}\right) + C_\eta \exp(-v_\eta n)$$
$$\leq \Pr\left(\sum_{ij} u_{ij}A_{ij} > c\beta_l\varphi_n/2\right) + C_\eta \exp(-v_\eta n)$$
$$\leq \exp(-C_c n\sqrt{n\rho_n}) + C_\eta \exp(-v_\eta n)$$

$$\leq C_{c,\eta}\exp(-v_{c,\eta}n),$$

where $C_{c,\eta}$ and $v_{c,\eta}$ are two constants determined by c and η.

By a standard volume argument we have $|T| \leq e^{n\log(7/\delta)}$ (see Claim 2.9 of [58]), so the desired result follows from the union bound. □

C. Bounding the heavy pairs

By the same argument as appendix B.3 of [91], to bound $\sup_{\mathbf{x},\mathbf{y}\in\mathcal{T}} |\sum_{(i,j)\in\mathbb{H}(x,y)} x_i y_j w_{ij}|$, it suffices to show

$$\sum_{(i,j)\in\mathbb{H}} x_i y_j A'_{ij} = O(\sqrt{\varphi_n})$$

with high probability. Since $A'_{ij} = A_{ij}e^{-\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0} - (A_{ij}e^{-\mathbf{z}_{ij}^T\boldsymbol{\gamma}^0} - A_{ij}e^{-\mathbf{z}_{ij}^T\hat{\boldsymbol{\gamma}}}) \leq A_{ij}\beta_l^{-1}(1+o(1))$, we only need to show

$$\sum_{(i,j)\in\mathbb{H}} x_i y_j A_{ij} = O(\sqrt{\varphi_n})$$

with high probability. Define $A''_{ij} = \max(A_{ij}, 1)$, then $A''_{ij} \sim B(1, 1-e^{-\lambda_{ij}})$ and $A''_{ij} = A_{ij}(1-(1-A_{ij}^{-1})_+) = A_{ij}(1+o(1))$, so it's sufficient to show

$$\sum_{(i,j)\in\mathbb{H}} x_i y_j A''_{ij} = O(\sqrt{\varphi_n}).$$

Because $\sum_j(1-e^{-\lambda_{ij}}) = O(\sqrt{\varphi_n})$, so the problem is exactly equivalent to [91]. We can directly get the following lemma.

Lemma 3.15. *(Heavy pair bound). For any given $r > 0$, there exists a constant C_r such that*

$$\sup_{x,y\in T} |\sum_{(i,j)\in\mathbb{H}} x_i y_j w_{ij}| \leq C_r\sqrt{\varphi_n}$$

with probability at least $1-2n^{-r}$.

Conclusion

In summary, we propose two extensions of vanilla SBM. One is by adding more layers with the same community membership. By studying the asymptotic distribution of eigenvalues and eigenvectors, we found the explicit formula for the optimal weight, and monotonic relation between eigenvalue gap and SNR. Based on above findings, we develop two algorithms to minimize spectral clustering error, which both outperform other state-of-art models. The other modify the SBM by incorporating pairwise covariates into the edge generation procedure to create more flexibility for the original model. The consistency property of MLE is proved and an efficient algorithm SCWA is proposed specifically to estimate the model. Our algorithms show potentials for involving pairwise covariates in both simulations and real data.

Here we list a few possible future research directions following this thesis.

- In Chapter 2, we only consider assortative matrices, which automatically results in positive weights. If we take disassortative matrices into account, one natural idea is to use negative weights. However, the behavior of eigenvalues and eigenvectors would change due to the negative weights, and that would be an interesting topic to deal with.

- We focus our theorem on balanced MPPM, which is kind of limited in practice. We may consider whether the eigenvalue gap still works under more general settings, say MSBM or unbalanced model, or more complicated quantities should come into the picture. It also remains to be a question whether there is an explicit formula under more general settings.

- In Chapter 3, we discussed the covariates for a fixed dimension p, it would be natural to ask the theoretical property when *p* grows with *n*, just like the high-dimensional regression.

- In our assumption, we require *Z* to be independent since some concentration inequalities may not hold under dependent situations. If we construct pairwise covariates from nodal covariates, the independent assumption will not hold. In this case, whether we could still use PCABM would be an interesting research topic.

References

[1] E. Abbe, A. S. Bandeira, and G. Hall, "Exact recovery in the stochastic block model," *IEEE Transactions on Information Theory*, vol. 62, no. 1, pp. 471–487, 2015.

[2] E. Abbe and C. Sandon, "Achieving the ks threshold in the general stochastic block model with linearized acyclic belief propagation," in *Advances in Neural Information Processing Systems*, 2016, pp. 1334–1342.

[3] ——, "Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, IEEE, 2015, pp. 670–688.

[4] ——, "Proof of the achievability conjectures for the general stochastic block model," *Communications on Pure and Applied Mathematics*, vol. 71, no. 7, pp. 1334–1406, 2018.

[5] ——, "Recovering communities in the general stochastic block model without knowing the parameters," in *Advances in neural information processing systems*, 2015, pp. 676–684.

[6] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: Divided they blog," in *Proceedings of the 3rd international workshop on Link discovery*, ACM, 2005, pp. 36–43.

[7] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership analysis of high-throughput interaction studies: Relational data," *arXiv preprint arXiv:0706.0294*, 2007.

[8] ——, "Mixed membership stochastic blockmodels," *Journal of machine learning research*, vol. 9, no. Sep, pp. 1981–2014, 2008.

[9] E. Airoldi, D. Blei, E. Xing, and S. Fienberg, "A latent mixed membership model for relational data," in *Proceedings of the 3rd international workshop on Link discovery*, 2005, pp. 82–89.

[10] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.

[11] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "Np-hardness of euclidean sum-of-squares clustering," *Machine learning*, vol. 75, no. 2, pp. 245–248, 2009.

[12] A. A. Amini, A. Chen, P. J. Bickel, E. Levina, *et al.*, "Pseudo-likelihood methods for community detection in large sparse networks," *The Annals of Statistics*, vol. 41, no. 4, pp. 2097–2122, 2013.

[13] A. A. Amini, E. Levina, *et al.*, "On semidefinite relaxations for the block model," *The Annals of Statistics*, vol. 46, no. 1, pp. 149–179, 2018.

[14] P. Arabie, S. A. Boorman, and P. R. Levitt, "Constructing blockmodels: How and why," *Journal of mathematical psychology*, vol. 17, no. 1, pp. 21–63, 1978.

[15] K. Avrachenkov, L. Cottatellucci, and A. Kadavankandy, "Spectral properties of random matrices for stochastic block model," in *2015 13th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, IEEE, 2015, pp. 537–544.

[16] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh, "Noise thresholds for spectral clustering," in *Advances in Neural Information Processing Systems*, 2011, pp. 954–962.

[17] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, "Information-theoretic thresholds for community detection in sparse networks," in *Conference on Learning Theory*, 2016, pp. 383–416.

[18] J. Baumes, M. Goldberg, M. Magdon-Ismail, and W. Al Wallace, "Discovering hidden groups in communication networks," in *International Conference on Intelligence and Security Informatics*, Springer, 2004, pp. 378–389.

[19] B. BBall and M. EJNewman, "Anefficientandprincipled methodfordetectingcommunitiesinnetworks," *Physical ReviewE*, vol. 84, p. 036 103, 2011.

[20] J. E. Beasley, "Heuristic algorithms for the unconstrained binary quadratic programming problem," *London, UK: Management School, Imperial College*, vol. 4, 1998.

[21] F. Benaych-Georges, C. Bordenave, and A. Knowles, "Spectral radii of sparse random matrices," *arXiv preprint arXiv:1704.02945*, 2017.

[22] F. Benaych-Georges, C. Bordenave, A. Knowles, *et al.*, "Largest eigenvalues of sparse inhomogeneous erdős–rényi graphs," *The Annals of Probability*, vol. 47, no. 3, pp. 1653–1676, 2019.

[23] F. Benaych-Georges and R. R. Nadakuditi, "The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices," *Advances in Mathematics*, vol. 227, no. 1, pp. 494–521, 2011.

[24] P. J. Bickel and A. Chen, "A nonparametric view of network models and newman–girvan and other modularities," *Proceedings of the National Academy of Sciences*, vol. 106, no. 50, pp. 21 068–21 073, 2009.

[25] P. J. Bickel and P. Sarkar, "Hypothesis testing for automated community detection in networks," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 1, pp. 253–273, 2016.

[26] P. Bickel, D. Choi, X. Chang, and H. Zhang, "Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels," *Ann. Statist.*, vol. 41, no. 4, pp. 1922–1943, Aug. 2013.

[27] N. Binkiewicz, J. T. Vogelstein, and K. Rohe, "Covariate-assisted spectral clustering," *Biometrika*, vol. 104, no. 2, pp. 361–377, 2017.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[29] B. Bollobás, S. Janson, and O. Riordan, "The phase transition in inhomogeneous random graphs," *Random Structures & Algorithms*, vol. 31, no. 1, pp. 3–122, 2007.

[30] R. B. Boppana, "Eigenvalues and graph bisection: An average-case analysis," in *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, IEEE, 1987, pp. 280–285.

[31] C. Bordenave, M. Lelarge, and L. Massoulié, "Non-backtracking spectrum of random graphs: Community detection and non-regular ramanujan graphs," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, IEEE, 2015, pp. 1347–1357.

[32] U. Braun, A. Schäfer, H. Walter, S. Erk, N. Romanczuk-Seiferth, L. Haddad, J. I. Schweiger, O. Grimm, A. Heinz, H. Tost, *et al.*, "Dynamic reconfiguration of frontal brain networks during executive cognition in humans," *Proceedings of the National Academy of Sciences*, vol. 112, no. 37, pp. 11 678–11 683, 2015.

[33] T. N. Bui, S. Chaudhuri, F. T. Leighton, and M. Sipser, "Graph bisection algorithms with good average case behavior," *Combinatorica*, vol. 7, no. 2, pp. 171–191, 1987.

[34] T. T. Cai, X. Li, *et al.*, "Robust and computationally feasible community detection in the presence of arbitrary outlier nodes," *The Annals of Statistics*, vol. 43, no. 3, pp. 1027–1059, 2015.

[35] A. Celisse, J.-J. Daudin, L. Pierre, *et al.*, "Consistency of maximum-likelihood and variational estimators in the stochastic block model," *Electronic Journal of Statistics*, vol. 6, pp. 1847–1899, 2012.

[36] K. Chaudhuri, F. Chung, and A. Tsiatas, “Spectral clustering of graphs with general degrees in the extended planted partition model,” in *Conference on Learning Theory*, 2012, pp. 35–1.

[37] A. Chen, A. A. Amini, E. Levina, and P. J. Bickel, “Fitting community models to large sparse networks,” *Ann. Stat.*, vol. 41, no. arXiv: 1207.2340, pp. 2097–2122, 2012.

[38] J. Chen and B. Yuan, “Detecting functional modules in the yeast protein–protein interaction network,” *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[39] K. Chen and J. Lei, “Network cross-validation for determining the number of communities in network data,” *Journal of the American Statistical Association*, 2016, just-accepted.

[40] P.-Y. Chen and A. O. Hero, “Multilayer spectral graph clustering via convex layer aggregation: Theory and algorithms,” *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3, no. 3, pp. 553–567, 2017.

[41] Y. Chen, S. Sanghavi, and H. Xu, “Clustering sparse graphs,” in *Advances in neural information processing systems*, 2012, pp. 2204–2212.

[42] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, “Stochastic blockmodels with a growing number of classes,” *Biometrika*, asr053, 2012.

[43] F. Chung, F. R. Chung, F. C. Graham, L. Lu, K. F. Chung, *et al.*, *Complex graphs and networks*, 107. American Mathematical Soc., 2006.

[44] A. Clauset, “Finding local community structure in networks,” *Physical review E*, vol. 72, no. 2, p. 026 132, 2005.

[45] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, *et al.*, “Integration of biological networks and gene expression data using cytoscape,” *Nature protocols*, vol. 2, no. 10, p. 2366, 2007.

[46] A. Condon and R. M. Karp, “Algorithms for graph partitioning on the planted partition model,” in *Randomization, Approximation, and Combinatorial Optimization. Algorithms and Techniques*, Springer, 1999, pp. 221–232.

[47] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, “Comparing community structure identification,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, P09008, 2005.

[48] J.-J. Daudin, F. Picard, and S. Robin, “A mixture model for random graphs,” *Statistics and computing*, vol. 18, no. 2, pp. 173–183, 2008.

[49] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications," *Physical Review E*, vol. 84, no. 6, p. 066 106, 2011.

[50] P. Doreian, V. Batagelj, and A. Ferligoj, *Generalized blockmodeling*. Cambridge university press, 2005, vol. 25.

[51] P. Doreian, V. Batagelj, and A. Ferligoj, "Generalized blockmodeling of two-mode network data," *Social networks*, vol. 26, no. 1, pp. 29–53, 2004.

[52] R. Durrett, *Random graph dynamics*, 7. Cambridge university press Cambridge, 2007, vol. 200.

[53] M. E. Dyer and A. M. Frieze, "The solution of some random np-hard problems in polynomial expected time," *Journal of Algorithms*, vol. 10, no. 4, pp. 451–489, 1989.

[54] L. Erdős, H.-T. Yau, and J. Yin, "Rigidity of eigenvalues of generalized wigner matrices," *Advances in Mathematics*, vol. 229, no. 3, pp. 1435–1515, 2012.

[55] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publ. Math. Inst. Hung. Acad. Sci*, vol. 5, no. 1, pp. 17–60, 1960.

[56] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5220–5227, 2004.

[57] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *ACM SIGCOMM computer communication review*, vol. 29, no. 4, pp. 251–262, 1999.

[58] U. Feige and E. Ofek, "Spectral techniques applied to sparse random graphs," *Random Structures & Algorithms*, vol. 27, no. 2, pp. 251–275, 2005.

[59] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, "Statistical analysis of multiple sociometric relations," *Journal of the american Statistical association*, vol. 80, no. 389, pp. 51–67, 1985.

[60] S. E. Fienberg and S. Wasserman, "An exponential family of probability distributions for directed graphs: Comment," *Journal of the American Statistical Association*, vol. 76, no. 373, pp. 54–57, 1981.

[61] D. E. Fishkind, D. L. Sussman, M. Tang, J. T. Vogelstein, and C. E. Priebe, "Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown," *SIAM Journal on Matrix Analysis and Applications*, vol. 34, no. 1, pp. 23–39, 2013.

[62] G. W. Flake, S. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of web communities," *Computer*, vol. 35, no. 3, pp. 66–70, 2002.

[63] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[64] O. Frank and D. Strauss, "Markov graphs," *Journal of the american Statistical association*, vol. 81, no. 395, pp. 832–842, 1986.

[65] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, "Achieving optimal misclassification proportion in stochastic block models," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 1980–2024, 2017.

[66] F. Glover and M. Laguna, "Tabu search," in *Handbook of combinatorial optimization*, Springer, 1998, pp. 2093–2229.

[67] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi, "A survey of statistical network models," *Foundations and Trends® in Machine Learning*, vol. 2, no. 2, pp. 129–233, 2010.

[68] S. Greenland, J. M. Robins, and J. Pearl, "Confounding and collapsibility in causal inference," *Statistical science*, pp. 29–46, 1999.

[69] Q. Han, K. Xu, and E. Airoldi, "Consistent estimation of dynamic and multi-layer block models," in *International Conference on Machine Learning*, 2015, pp. 1511–1520.

[70] M. S. Handcock, A. E. Raftery, and J. M. Tantrum, "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, no. 2, pp. 301–354, 2007.

[71] S. Hill, F. Provost, C. Volinsky, *et al.*, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science*, vol. 21, no. 2, pp. 256–276, 2006.

[72] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the american Statistical association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[73] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social networks*, vol. 5, no. 2, pp. 109–137, 1983.

[74] P. W. Holland and S. Leinhardt, "An exponential family of probability distributions for directed graphs," *Journal of the american Statistical association*, vol. 76, no. 373, pp. 33–50, 1981.

[75] J. Hu, H. Qin, T. Yan, and Y. Zhao, "On consistency of model selection for stochastic block models," *arXiv preprint arXiv:1611.01238*, 2016.

[76] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, no. 1, pp. 193–218, 1985.

[77] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on knowledge and data engineering*, vol. 16, no. 11, pp. 1370–1386, 2004.

[78] J. Jin, "Fast community detection by score," *The Annals of Statistics*, vol. 43, no. 1, pp. 57–89, 2015.

[79] G. Karlebach and R. Shamir, "Modelling and analysis of gene regulatory networks," *Nature Reviews Molecular Cell Biology*, vol. 9, no. 10, pp. 770–780, 2008.

[80] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, p. 016 107, 2011.

[81] A. Khorunzhy, "Sparse random matrices: Spectral edge and statistics of rooted trees," *Advances in Applied Probability*, vol. 33, no. 1, pp. 124–140, 2001.

[82] J. Kleinberg, "The small-world phenomenon: An algorithmic perspective," in *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 2000, pp. 163–170.

[83] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, no. 6798, pp. 845–845, 2000.

[84] ——, "Small-world phenomena and the dynamics of information," in *Advances in neural information processing systems*, 2002, pp. 431–438.

[85] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, "Spectral redemption in clustering sparse networks," *Proceedings of the National Academy of Sciences*, vol. 110, no. 52, pp. 20 935–20 940, 2013.

[86] A. Kumar, Y. Sabharwal, and S. Sen, "A simple linear time (1+/spl epsiv/)-approximation algorithm for k-means clustering in any dimensions," in *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*, IEEE, 2004, pp. 454–462.

[87] C. M. Le and E. Levina, "Estimating the number of communities in networks by spectral methods," *arXiv preprint arXiv:1507.00827*, 2015.

[88] C. M. Le, E. Levina, and R. Vershynin, "Concentration and regularization of random graphs," *Random Structures & Algorithms*, vol. 51, no. 3, pp. 538–561, 2017.

[89] ——, “Sparse random graphs: Regularization and concentration of the laplacian,” *arXiv preprint arXiv:1502.03049*, 2015.

[90] J. Lei, “A goodness-of-fit test for stochastic block models,” *The Annals of Statistics*, vol. 44, no. 1, pp. 401–424, 2016.

[91] J. Lei and A. Rinaldo, “Consistency of spectral clustering in stochastic block models,” *The Annals of Statistics*, vol. 43, no. 1, pp. 215–237, 2015.

[92] T. Li, E. Levina, and J. Zhu, “Network cross-validation by edge sampling,” *arXiv preprint arXiv:1612.04717*, 2016.

[93] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[94] G. Linden, B. Smith, and J. York, “Amazon. com recommendations: Item-to-item collaborative filtering,” *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[95] F. Lorrain and H. C. White, “Structural equivalence of individuals in social networks,” *The Journal of mathematical sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[96] Y. Lu and H. H. Zhou, “Statistical and computational guarantees of lloyd’s algorithm and its variants,” *arXiv preprint arXiv:1612.02099*, 2016.

[97] V. Lyzinski, D. L. Sussman, M. Tang, A. Athreya, C. E. Priebe, *et al.*, “Perfect clustering for stochastic blockmodel graphs via adjacency spectral embedding,” *Electronic Journal of Statistics*, vol. 8, no. 2, pp. 2905–2922, 2014.

[98] Z. Ma and Z. Ma, “Exploration of large networks via fast and universal latent space model fitting,” *arXiv preprint arXiv:1705.02372*, 2017.

[99] J. R. Magnus, “On differentiating eigenvalues and eigenvectors,” *Econometric Theory*, vol. 1, no. 2, pp. 179–191, 1985.

[100] E. M. Marcotte, M. Pellegrini, H.-L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg, “Detecting protein function and protein-protein interactions from genome sequences,” *Science*, vol. 285, no. 5428, pp. 751–753, 1999.

[101] L. Massoulié, “Community detection thresholds and the weak ramanujan property,” in *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, 2014, pp. 694–703.

[102] F. McSherry, “Spectral partitioning of random graphs,” in *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, IEEE, 2001, pp. 529–537.

[103] S. Milgram, “The small world problem,” *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[104] J. L. Moreno, “Who shall survive?: A new approach to the problem of human interrelations.,” 1934.

[105] E. Mossel, J. Neeman, and A. Sly, “Consistency thresholds for binary symmetric block models,” *arXiv preprint arXiv:1407.1591*, vol. 3, no. 5, 2014.

[106] ——, “Reconstruction and estimation in the planted partition model,” *Probability Theory and Related Fields*, vol. 162, no. 3-4, pp. 431–461, 2015.

[107] M. E. Newman, “Detecting community structure in networks,” *The European Physical Journal B*, vol. 38, no. 2, pp. 321–330, 2004.

[108] ——, “Finding community structure in networks using the eigenvectors of matrices,” *Physical review E*, vol. 74, no. 3, p. 036 104, 2006.

[109] ——, “Modularity and community structure in networks,” *Proceedings of the national academy of sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[110] ——, “The structure and function of complex networks,” *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003.

[111] M. E. Newman and A. Clauset, “Structure and inference in annotated networks,” *Nature communications*, vol. 7, 2016.

[112] K. Nowicki and T. A. B. Snijders, “Estimation and prediction for stochastic blockstructures,” *Journal of the American statistical association*, vol. 96, no. 455, pp. 1077–1087, 2001.

[113] P. Pattison and S. Wasserman, “Logit models and logistic regressions for social networks: Ii. multivariate relations,” *British Journal of Mathematical and Statistical Psychology*, vol. 52, no. 2, pp. 169–193, 1999.

[114] S. Paul and Y. Chen, “Consistency of community detection in multi-layer networks using spectral and matrix factorization methods,” *arXiv preprint arXiv:1704.07353*, 2017.

[115] L. Peel, D. B. Larremore, and A. Clauset, “The ground truth about metadata and community detection in networks,” *Science Advances*, vol. 3, no. 5, e1602548, 2017.

[116] X. Qi, W. Tang, Y. Wu, G. Guo, E. Fuller, and C.-Q. Zhang, “Optimal local community detection in social networks based on density drop of subgraphs,” *Pattern Recognition Letters*, vol. 36, pp. 46–53, 2014.

[117] T. Qin and K. Rohe, "Regularized spectral clustering under the degree-corrected stochastic blockmodel," in *Advances in Neural Information Processing Systems*, 2013, pp. 3120–3128.

[118] G. Robins, P. Pattison, and S. Wasserman, "Logit models and logistic regressions for social networks: Iii. valued relations," *Psychometrika*, vol. 64, no. 3, pp. 371–394, 1999.

[119] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *The Annals of Statistics*, pp. 1878–1915, 2011.

[120] S. Sahebi and W. W. Cohen, "Community-based recommendations: A solution to the cold start problem," in *Workshop on recommender systems and the social web, RSWEB*, 2011, p. 60.

[121] D. F. Saldana, Y. Yu, and Y. Feng, "How many communities are there?" *Journal of Computational and Graphical Statistics*, vol. 26, no. 1, pp. 171–181, 2017.

[122] P. Sarkar, P. J. Bickel, *et al.*, "Role of normalization in spectral clustering for stochastic blockmodels," *The Annals of Statistics*, vol. 43, no. 3, pp. 962–990, 2015.

[123] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[124] G. Simmel, *The sociology of georg simmel*. Simon and Schuster, 1950, vol. 92892.

[125] T. A. Snijders and K. Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of classification*, vol. 14, no. 1, pp. 75–100, 1997.

[126] D. Strauss and M. Ikeda, "Pseudolikelihood estimation for social networks," *Journal of the American statistical association*, vol. 85, no. 409, pp. 204–212, 1990.

[127] L. S. Tan, A. H. Chan, and T. Zheng, "Topic-adjusted visibility metric for scientific articles," *The Annals of Applied Statistics*, vol. 10, no. 1, pp. 1–31, 2016.

[128] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, Y. Park, and C. E. Priebe, "A semiparametric two-sample hypothesis testing problem for random graphs," *Journal of Computational and Graphical Statistics*, vol. 26, no. 2, pp. 344–354, 2017.

[129] M. Tang, C. E. Priebe, *et al.*, "Limit theorems for eigenvectors of the normalized laplacian for random graphs," *The Annals of Statistics*, vol. 46, no. 5, pp. 2360–2415, 2018.

[130] J. Travers and S. Milgram, "An experimental study of the small world problem," in *Social Networks*, Elsevier, 1977, pp. 179–197.

[131] T. Van Laarhoven and E. Marchiori, “Local network community detection with continuous optimization of conductance and weighted kernel k-means,” *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 5148–5175, 2016.

[132] U. Von Luxburg, “A tutorial on spectral clustering,” *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[133] V. Vu, “A simple svd algorithm for finding hidden partitions,” *Combinatorics, Probability and Computing*, vol. 27, no. 1, pp. 124–140, 2018.

[134] Y. R. Wang, P. J. Bickel, *et al.*, “Likelihood-based model selection for stochastic block models,” *The Annals of Statistics*, vol. 45, no. 2, pp. 500–528, 2017.

[135] S. Wasserman, K. Faust, *et al.*, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[136] S. Wasserman and P. Pattison, “Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp,” *Psychometrika*, vol. 61, no. 3, pp. 401–425, 1996.

[137] Y.-C. Wei and C.-K. Cheng, “Towards efficient hierarchical designs by ratio cut partitioning,” in *1989 IEEE International Conference on Computer-Aided Design. Digest of Technical Papers*, IEEE, 1989, pp. 298–301.

[138] J. A. Wellner, “Empirical processes: Theory and applications,” *Notes for a course given at Delft University of Technology*, 2005.

[139] H. Weng and Y. Feng, “Community detection with nodal information,” *arXiv preprint arXiv:1610.09735*, 2016.

[140] R. Wu, J. Xu, R. Srikant, L. Massoulié, M. Lelarge, and B. Hajek, “Clustering and inference from pairwise comparisons,” in *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, 2015, pp. 449–450.

[141] Y. Wu, R. Jin, J. Li, and X. Zhang, “Robust local community detection: On free rider effect and its elimination,” *Proceedings of the VLDB Endowment*, vol. 8, no. 7, pp. 798–809, 2015.

[142] Y.-J. Wu, E. Levina, and J. Zhu, “Generalized linear models with low rank effects for network data,” *arXiv preprint arXiv:1705.06772*, 2017.

[143] B. Yan, P. Sarkar, and X. Cheng, “Exact recovery of number of blocks in blockmodels,” *arXiv preprint arXiv:1705.08580*, 2017.

[144] T. Yan, B. Jiang, S. E. Fienberg, and C. Leng, “Statistical inference in a directed network model with covariates,” *Journal of the American Statistical Association*, vol. 114, no. 526, pp. 857–868, 2019.

[145] Y. Zhang, E. Levina, and J. Zhu, “Community detection in networks with node features,” *Electronic Journal of Statistics*, vol. 10, no. 2, pp. 3153–3178, 2016.

[146] Y. Zhao, E. Levina, and J. Zhu, “Community extraction for social networks,” *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7321–7326, 2011.

[147] ——, “Consistency of community detection in networks under degree-corrected stochastic block models,” *The Annals of Statistics*, vol. 40, no. 4, pp. 2266–2292, 2012.